THÈSE DE DOCTORAT DE MATHÉMATIQUES
de l'Université Blaise Pascal (Clermont-Ferrand 2)
préparée au Laboratoire de Mathématiques
UMR 6620 - CNRS
École Doctorale des Sciences Fondamentales N° 733

N° d'ordre : D. U. 2300

Thèse de doctorat

Spécialité Mathématiques Appliquées

présentée par

Marianne Bessemoulin-Chatard

Développement et analyse de schémas volumes finis motivés par la préservation de comportements asymptotiques. Application à des modèles issus de la physique et de la biologie

Soutenue publiquement le 30 novembre 2012

Après avis de :

Franck	Boyer	Aix-Marseille Université
Mario	Ohlberger	Universität Münster

Devant le jury composé de :

Christophe	Berthon	Université de Nantes	Examinateur
Franck	Boyer	Aix-Marseille Université	Rapporteur
Claire	Chainais-Hillairet	Université Lille 1	Directrice de thèse
Francis	Filbet	Université Lyon 1	Directeur de thèse
Pauline	Lafitte-Godillon	École Centrale Paris	Examinatrice
Yue-Jun	Peng	Université Blaise Pascal	Examinateur

À Stéphane et Léo

Remerciements

Mes premiers remerciements s'adressent naturellement à mes deux directeurs de thèse, Claire Chainais-Hillairet et Francis Filbet. Ils m'ont initiée à la recherche en me proposant des sujets passionnants, et en étant toujours disponibles pour répondre à mes questions et me conseiller. Je les remercie également de m'avoir fait rencontrer autant de personnes intéressantes et de m'avoir permis de participer à de nombreuses conférences.

Je remercie chaleureusement Franck Boyer et Mario Ohlberger pour avoir accepté d'être les rapporteurs de cette thèse et pour leurs remarques judicieuses.
Je tiens aussi à remercier sincèrement Christophe Berthon, Pauline Lafitte-Godillon et Yue-Jun Peng pour leur présence au sein de mon jury.

Je tiens également à remercier les différents collaborateurs avec qui j'ai eu la chance de travailler durant cette thèse : merci encore une fois à Claire Chainais-Hillairet et Francis Filbet, à Ansgar Jüngel ainsi qu'à Marie-Hélène Vignal. J'espère avoir l'occasion de travailler à nouveau avec eux à l'avenir.

J'adresse un grand merci au Laboratoire de Mathématiques de Clermont-Ferrand. Merci en particulier à Marie-Paule Bressoulaly et Séverine Miginiac pour leur gentillesse et leur disponibilité. Je remercie également sincèrement Karine Darot et Valérie Sourlier pour toute l'aide qu'elles m'ont apportée lors de l'organisation de mes déplacements.
Je remercie aussi l'Institut Camille Jordan de Lyon où j'ai passé pas mal de temps durant ces trois années.
Enfin, je remercie chaleureusement l'Institut d'Analyse et de Calcul Scientifique de l'Université Technologique de Vienne, où j'ai eu l'occasion de faire un séjour de trois mois durant cette thèse.

J'en arrive maintenant aux remerciements pour mes amis doctorants ou jeunes docteurs qu'il m'a été donné de rencontrer et de fréquenter au cours de ces trois années de thèse.
Je pense tout d'abord à mes trois demi-frère et sœurs de thèse. Valérie, tu achevais ta thèse alors que je la commençais, nous n'avons donc pas pu nous voir très longtemps, mais je garde un très bon souvenir de ta visite à Vienne. Amélie, tu me précédais d'un an et tu as toujours su m'encourager, du début à la fin de cette thèse. Tu as vraiment joué un rôle de « grande sœur »pour moi ! Enfin Thomas, mon « jumeau », on en aura passé du temps à discuter, que ce soit de maths ou de toute autre chose. Je te souhaite le meilleur pour la suite, ainsi qu'à Fred. J'espère que vous passerez un super séjour aux États-Unis, mais que vous reviendrez quand même nous voir de temps en temps !
Je tiens ensuite à remercier mes collègues du bureau 111B, grâce à qui j'ai pu faire cette thèse dans les meilleures conditions : Erwan, Romain et Cécile, qui rédigeaient en même

temps que moi et contribuaient à créer une ambiance studieuse (ou pas...), Julien et ses histoires improbables qui sait toujours mettre l'ambiance, Emmanuelle, Mohamed, et les « petits nouveaux »Adriane et Raouf.
J'ai aussi passé une partie de ma thèse dans le bureau 1213, dont je tiens également à remercier les occupants : Mehdi, Stéphane et Xinyu.
Merci aussi à tous ceux de Clermont, de Lyon ou d'ailleurs avec qui j'ai passé du temps, autour d'un café, d'un thé ou d'une bière : Alina, Bérénice, Blanche, Claire, Emna, Fanny, Flore, Franck, Fred et Sophie, Gaëlle, Hugo, Ihsane, J.-B., Joël, Jonathan, Manon, Mickaël, Mohamed, Rudy, Thomas.
Je remercie aussi chaleureusement mes collègues de Vienne qui m'ont accueillie si gentiment : Birgit, Dominik, Elan, Inès, Jan-Frederik, Mario, Sabine.

C'est maintenant au tour de mes amis « non matheux » : Alexia et Anthony, Annabelle, Anne et Antoine, Bastien et Laure, Caroline et Éric, Erwin, Florie et Jean-Michel, Hélène, Hugo et Anne-Laure, Kevin, Laurie, Nicolas et Sylvie, Thomas. Heureusement que vous étiez là pour me rappeler qu'il y avait une vie en dehors de la thèse ! J'ai également une grosse pensée pour toute la Sama Family.

Je tiens enfin à remercier ceux qui ont toujours été là : ma famille. Un grand merci tout d'abord à mes parents, Brigitte et Bruno, qui m'ont toujours soutenue dans mes choix ; sans eux je n'en serais pas là aujourd'hui. Je pense également à mes grand-parents pour tout l'amour qu'ils m'ont donné. Je remercie très fort mes frère et sœurs, Camille, Claude et Delphine. Mention spéciale à Delphine pour toute l'aide qu'elle m'a apportée en me relisant et en me coachant en anglais. Je tiens également à remercier mes beaux-parents, Chantal et Jacky.
Je conclurais ces remerciements par celui sans qui ce manuscrit n'existerait pas aujourd'hui, Stéphane. Merci pour ton amour, ta compréhension, ton soutien, et surtout ton infinie patience. Enfin je pense à notre fils Léo, qui m'a accompagnée pendant toute la phase finale de rédaction de cette thèse, et qui est maintenant chaque jour une source de bonheur pour nous.

Table des matières

Table des figures

Liste des tableaux

Avant-propos

Nous nous intéressons dans cette thèse au développement et à l'analyse de schémas numériques de type volumes finis pour des équations de convection-diffusion, qui apparaissent notamment dans des modèles issus de la physique ou de la biologie. Ce manuscrit s'articule en trois parties :

- Dans la première partie, nous considèrons la discrétisation du système de dérive-diffusion linéaire pour les semi-conducteurs par le schéma de Scharfetter-Gummel implicite. Nous nous intéressons à la préservation par ce schéma de deux types d'asymptotiques (l'asymptotique en temps long et la limite quasi-neutre) en prouvant des estimations d'énergie–dissipation d'énergie discrètes.
- Dans la seconde partie, nous nous intéressons à des schémas volumes finis préservant l'asymptotique en temps long dans un cadre plus général. Plus précisément, nous considèrons des équations de type convection-diffusion non linéaires, qui apparaissent dans plusieurs contextes physiques : équation des milieux poreux, système de dérive-diffusion pour les semi-conducteurs... Nous proposons deux discrétisations en espace permettant de préserver le comportement en temps long des solutions approchées.
- Dans la troisième partie, nous étudions un schéma numérique pour un modèle de chimiotactisme avec diffusion croisée. L'étude de la convergence du schéma repose sur une estimation d'entropie qui fait intervenir des inégalités fonctionnelles discrètes de type Poincaré-Sobolev et Gagliardo-Nirenberg-Sobolev, et dont la démonstration fait l'objet d'un chapitre indépendant.

Les différents chapitres sont composés des travaux suivants :

- Les chapitres 1 et 2 regroupent un travail publié dans les *proceedings* du congrès FVCA VI, 6th International Symposium on Finite Volume for Complex Applications, sous le titre *Asymptotic Behavior of the Scharfetter–Gummel Scheme for the Drift-Diffusion Model* [55], et des résultats récents obtenus en collaboration avec Claire Chainais-Hillairet et Marie-Hélène Vignal, *Convergence of a fully implicit scheme for the drift-diffusion system. Stability at the quasineutral limit* [18].
- Le chapitre 3 est un article accepté pour publication dans Numerische Mathematik, *A finite volume scheme for convection-diffusion equations with nonlinear diffusion derived from the Scharfetter-Gummel scheme* [16].
- Le chapitre 4 est un travail réalisé en collaboration avec Francis Filbet, *A finite volume scheme for nonlinear degenerate parabolic equations* [19], accepté pour pub-

lication dans SIAM Journal on Scientific Computing.

- Le chapitre 5 est un article réalisé en collaboration avec Claire Chainais-Hillairet et Francis Filbet, *On discrete functional inequalities for some finite volume schemes* [17], soumis pour publication.
- Le chapitre 6 est le résultat d'une collaboration avec Ansgar Jüngel, suite à un séjour de trois mois à l'Université technique de Vienne. L'article en résultant, *A finite volume scheme for a Keller-Segel model with additional cross-diffusion* [20], est actuellement en préparation.

Introduction générale et présentation des travaux

Sommaire

Cette introduction est divisée en trois sections qui correspondent aux trois parties du manuscrit. Pour chacune de ces parties, nous commençons par introduire le cadre général et établir un bref état de l'art non exhaustif, avant de présenter les motivations des travaux effectués et les résultats obtenus.

1 Étude d'un schéma préservant des asymptotiques : le schéma de Scharfetter-Gummel tout implicite

La partie I de cette thèse concerne l'étude de la préservation d'asymptotiques par le schéma de Scharfetter-Gummel implicite pour le système de dérive-diffusion.
Nous commençons par introduire le modèle de dérive-diffusion en le resituant dans la hiérarchie de modèles existants pour la description des semi-conducteurs. Nous rappelons ensuite les résultats concernant le comportement en temps long de ce modèle ainsi que la limite quasi-neutre (limite lorsque la longueur de Debye tend vers zéro). Nous expliquons notamment le principe des preuves développées dans le contexte continu, qui

reposent sur des estimations d'énergie avec contrôle de la dissipation, puisque le point central de la partie I est l'adaptation de ces méthodes au niveau discret. Nous présentons ensuite un état de l'art synthétique concernant les discrétisations proposées pour le modèle de dérive-diffusion, incluant le schéma de Scharfetter-Gummel. Dans un troisième temps, nous résumons la motivation des travaux effectués dans cette partie en expliquant le principe général des schémas préservant des asymptotiques. Enfin, nous résumons les résultats obtenus dans cette partie I, à savoir : la convergence en temps long des solutions numériques obtenues par le schéma de Scharfetter-Gummel implicite vers une approximation de l'équilibre thermique (chapitre 1) et la convergence indépendamment de la valeur de la longueur de Debye des solutions approchées obtenues avec ce schéma vers une solution faible du système de dérive-diffusion quand le paramètre de discrétisation tend vers zéro (chapitre 2).

1.1 Le modèle de dérive-diffusion pour les semi-conducteurs

Hiérarchie de modèles

Il existe deux grandes catégories de modèles classiques pour les semi-conducteurs :
- les modèles cinétiques, qui se placent à l'échelle des électrons (par exemple, l'équation de Boltzmann pour les semi-conducteurs),
- les modèles fluides, qui assimilent l'ensemble des porteurs de charges à un fluide. Ces modèles décrivent directement des quantités macroscopiques telles que les densités de charge, la température, *etc.*

Si les modèles cinétiques sont plus précis pour prendre en compte des caractéristiques physiques, ils requièrent un coût élevé pour effectuer des simulations numériques. Ainsi, des modèles plus simples offrant un bon compromis entre précision physique et coût de calcul sont souvent considérés. Parmi ces modèles macroscopiques, une hiérarchie peut de nouveau être opérée, depuis les systèmes hydrodynamiques de type Euler-Poisson jusqu'aux systèmes de dérive-diffusion, en passant par les modèles de transport d'énergie. De nombreuses études ont de fait été menées sur les systèmes de type Euler-Poisson pour les plasmas ou les semi-conducteurs. Cependant, ces systèmes présentent plusieurs difficultés, en particulier du point de vue numérique. Dans certaines situations, ils peuvent être approchés par des équations plus simples, obtenues en faisant tendre vers zéro un petit paramètre apparaissant dans le système de départ. Ce petit paramètre peut par exemple être le temps de relaxation (*zero-relaxation-time limit*) ou la masse d'électron (*zero-electron-mass limit*). Une hiérarchie plus précise de ces différents modèles est notamment présentée dans les livres de P. Markowich, C. Ringhofer et C. Schmeiser [137] et de A. Jüngel [120].

Nous nous intéressons ici au modèle de dérive-diffusion standard ou isotherme, initialement proposé par W. Van Roosbroeck en 1950 [162]. Il s'agit d'un modèle simple dans lequel la température des électrons n'est plus une inconnue du problème.
Nous notons $\Omega \subset \mathbb{R}^d$ ($d = 1$, 2 ou 3) le domaine borné occupé par le semi-conducteur. La préconcentration en électrons et en trous est appelée dopage et notée C. Le système de dérive-diffusion est constitué de deux équations de continuité sur la densité d'électrons,

notée N, et sur la densité de trous, notée P, et d'une équation de Poisson sur le potentiel électrique, noté Ψ. Il s'écrit de la manière suivante :

$$\begin{cases} \partial_t N - \operatorname{div}(\nabla N - N\nabla\Psi) = 0 \\ \partial_t P - \operatorname{div}(\nabla P + P\nabla\Psi) = 0 \quad \text{dans } \Omega \times (0,T). \\ \lambda^2 \Delta\Psi = N - P - C \end{cases} \tag{1.1}$$

Ce système comprend un seul paramètre : la longueur de Debye adimensionnée λ, rapport de la longueur de Debye (qui mesure l'échelle typique des interactions électriques dans le semi-conducteur) par la longueur caractéristique du semi-conducteur.
Les conditions aux limites sont de deux types : Dirichlet pour les contacts ohmiques et Neumann homogènes pour les parties isolées. Ainsi, la frontière Γ de Ω est décomposée en deux parties disjointes, Γ^D et Γ^N, et en notant ν la normale à Γ extérieure à Ω, les conditions aux limites s'écrivent :

$$\begin{cases} N = N^D, \quad P = P^D, \Psi = \Psi^D, & \text{sur } \Gamma^D \times (0,T), \\ \nabla N \cdot \nu = \nabla P \cdot \nu = \nabla\Psi \cdot \nu = 0, & \text{sur } \Gamma^N \times (0,T). \end{cases} \tag{1.2}$$

Enfin, le système est complété par des conditions initiales sur les densités N et P :

$$N(x,0) = N_0(x), \quad P(x,0) = P_0(x), \quad x \in \Omega. \tag{1.3}$$

A titre d'exemple, un résultat d'existence et d'unicité de solutions faibles du système de dérive-diffusion standard a été obtenu par H. Gajewski [91] et par P. A. Markowich, C. A. Ringhofer et C. Schmeiser [137].
L'objet du chapitre 1 étant l'étude du comportement en temps long d'un schéma numérique pour ce modèle, nous commençons par rappeler ici les résultats obtenus dans le contexte continu.

Comportement en temps long

Le comportement en temps long du système de dérive-diffusion isotherme (1.1)–(1.3) a été étudié dans l'article de H. Gajewski et K. Gärtner [92] à l'aide d'une méthode d'entropie-dissipation. La fonctionnelle d'énergie considérée dans cet article avait déjà été utilisée par M. Mock [140] pour étudier l'asymptotique en temps long du système dans un cas particulier.
Nous commençons par expliquer les grandes lignes de la méthode d'entropie–dissipation employée, celle-ci apparaissant à plusieurs reprises dans ce manuscrit. A cette fin, nous nous référons à l'article de synthèse sur les méthodes d'entropie de A. Arnold et coll. [10].
Étant donnée une équation aux dérivées partielles d'inconnue u, la stratégie générale est la suivante :

- on commence par identifier l'état d'équilibre u_∞ et la fonctionnelle d'entropie E (ou plus généralement une fonction de Lyapunov associée à l'équation).
- la fonctionnelle E atteint son minimum à l'équilibre u_∞. L'écart entre une fonction u et l'équilibre u_∞ est mesuré par l'entropie relative

$$E[u|u_\infty] := E(u) - E(u_\infty).$$

On montre d'abord la convergence en entropie de la solution, c'est-à-dire

$$E[u(t)|u_\infty] \to 0 \text{ quand } t \to +\infty,$$

puis on peut espérer en déduire des résultats de convergence dans des espaces L^p, en utilisant des inégalités fonctionnelles de type Csiszar-Kullback.

– on étudie pour cela la fonctionnelle de dissipation d'entropie I, qui est la dérivée par rapport au temps de l'entropie :

$$I\left(u(t)\right) := -\frac{d}{dt} E\left(u(t)\right). \tag{1.4}$$

On arrive alors parfois à prouver une inégalité d'entropie–production d'entropie de la forme

$$I(u) \geq \Theta\left(E[u|u_\infty]\right), \tag{1.5}$$

où $\Theta : s \mapsto \Theta(s)$ est continue et strictement positive pour $s > 0$. Grâce à ce contrôle de l'entropie relative par la dissipation d'entropie (1.5), on obtient finalement, en utilisant la définition de la dissipation (1.4), une inégalité du type

$$-\frac{d}{dt} E[u|u_\infty] \geq \Theta\left(E[u|u_\infty]\right), \tag{1.6}$$

qui permet de montrer la convergence de $E[u|u_\infty]$ vers 0. De plus, si on connaît assez bien Θ, on peut calculer un taux de convergence explicite.

Historiquement, cette méthode a d'abord été employée dans le domaine des équations cinétiques (équations de Boltzmann ou de Landau par exemple), mais elle a ensuite été appliquée à une grande variété de problèmes : équations de type Fokker-Planck linéaires ou non linéaires (par exemple l'équation des milieux poreux), équations paraboliques non linéaires d'ordre 4 (par exemple l'équation des couches minces),... Pour un panorama plus complet de cette méthode et de ses domaines d'applications, on pourra se reporter à [10] et aux références citées dans cet article.

Revenons maintenant au système de dérive-diffusion (1.1)–(1.3). Dans [92], H. Gajewski et K. Gärtner prouvent que la solution de ce système converge vers l'état d'équilibre thermique quand $t \to \infty$, si les conditions au bord de Dirichlet sont à l'équilibre. L'équilibre thermique est un état stationnaire particulier pour lequel les courants d'électrons ($\nabla N - N\nabla\Psi$) et de trous ($\nabla P + P\nabla\Psi$) s'annulent. L'existence de cet équilibre est étudiée par P. Markowich, C. Ringhofer et C. Schmeiser dans [137, 136]. Si les conditions de Dirichlet sont à l'équilibre, c'est-à-dire si N^D, $P^D > 0$ et si

$$\log(N^D) - \Psi^D = \alpha_N, \quad \log(P^D) + \Psi^D = \alpha_P \quad \text{sur } \Gamma^D, \tag{1.7}$$

alors l'équilibre thermique est défini par

$$N^{eq}(x) = e^{\alpha_N + \Psi^{eq}(x)}, \quad P^{eq}(x) = e^{\alpha_P - \Psi^{eq}(x)}, \quad x \in \Omega, \tag{1.8}$$

où Ψ^{eq} est la solution du problème elliptique non linéaire suivant :

$$\begin{cases} \lambda^2 \Delta \Psi^{eq} = e^{\alpha_N + \Psi^{eq}} - e^{\alpha_P - \Psi^{eq}} - C \quad \text{sur } \Omega, \\ \Psi^{eq} = \Psi^D \text{ sur } \Gamma^D, \quad \nabla \Psi^{eq} \cdot \nu = 0 \text{ sur } \Gamma^N. \end{cases} \tag{1.9}$$

La preuve de la convergence vers l'équilibre thermique est basée sur une estimation d'énergie avec le contrôle de la dissipation d'énergie. Plus précisément, nous introduisons la fonctionnelle d'énergie relative $\mathcal{E}$, qui est la déviation de l'énergie totale (somme des énergies internes des densités d'électrons et de trous et de l'énergie due au potentiel électrique) par rapport à l'équilibre thermique :

$$\begin{aligned} \mathcal{E}(t) = &\int_\Omega (H(N(t)) - H(N^{eq}) - \log(N^{eq})(N(t) - N^{eq}))\, dx \\ &+ \int_\Omega (H(P(t)) - H(P^{eq}) - \log(P^{eq})(P(t) - P^{eq}))\, dx \\ &+ \frac{\lambda^2}{2} \int_\Omega |\nabla(\Psi(t) - \Psi^{eq})|^2\, dx, \end{aligned} \tag{1.10}$$

où $H(s) = \int_1^s \log(\tau)\, d\tau$, et la dissipation d'énergie :

$$\mathcal{I}(t) = \int_\Omega \left(N(t) \left| \nabla \left(\log(N(t)) - \Psi(t) \right) \right|^2 + P(t) \left| \nabla \left(\log(P(t)) + \Psi(t) \right) \right|^2 \right) dx. \tag{1.11}$$

Il est alors prouvé dans [92] que $\mathcal{E}$ est une fonction de Lyapunov vérifiant l'inégalité suivante :

$$0 \leq \mathcal{E}(t) + \int_0^t \mathcal{I}(\tau)\, d\tau \leq \mathcal{E}(0). \tag{1.12}$$

Partant de cette inégalité, les auteurs prouvent, à l'aide d'un contrôle de l'énergie relative par la dissipation de la forme (1.6), la décroissance exponentielle vers 0 de $\mathcal{E}$ ainsi que la convergence en temps long à un taux exponentiel de la solution de (1.1)–(1.3) vers l'équilibre thermique (1.8)–(1.9).

Limite quasi-neutre

Rappelons à présent les résultats relatifs au passage à la limite quasi-neutre dans le modèle de dérive-diffusion, c'est-à-dire la limite quand la longueur de Debye adimensionnée λ tend vers 0 dans le système (1.1)–(1.3). Physiquement, cela revient à considérer des structures de grande échelle par rapport à la longueur de Debye. Pour de telles structures, les mouvements des électrons ne sont plus distingués des mouvements des trous.

La limite quasi-neutre est un problème très étudié pour le modèle hydrodynamique et pour le modèle cinétique de Vlasov-Poisson. Citons par exemple l'article de S. Cordier et E. Grenier [59] portant sur la limite quasi-neutre dans le système d'Euler-Poisson en une dimension d'espace. Concernant le modèle de Vlasov-Poisson, des résultats partiels ont été

obtenus par Y. Brenier et E. Grenier dans [28].
Le modèle de dérive-diffusion étant un modèle hydrodynamique simplifié, il est naturel de s'intéresser au passage à la limite quasi-neutre dans ce système. De premiers résultats formels pour le modèle standard ont été obtenus par C. Ringhofer [154], avant d'être justifiés rigoureusement par I. Gasser, D. Levermore, P. Markowich et C. Schmeiser [95] dans le cas de conditions aux limites de Neumann homogènes. Par ailleurs, des résultats concernant le modèle de dérive-diffusion isentropique ont également été obtenus par I. Gasser [94] pour des conditions de Neumann homogènes ainsi que par A. Jüngel et Y. J. Peng [121] dans le cas de conditions aux limites mixtes et d'un dopage nul.
Dans ces articles, les preuves reposent sur une méthode d'entropie. Formellement, à la limite $\lambda \to 0$, le système de dérive-diffusion standard $(\mathcal{P}^\lambda)$ donné par (1.1)–(1.3) devient, dans le cas d'un dopage nul $C = 0$:

$$(\mathcal{P}^0) \qquad \left\{ \begin{array}{c} N = P, \\ \partial_t N - \Delta N = 0, \\ -\mathrm{div}(N \nabla \Psi) = 0. \end{array} \right.$$

Pour prouver rigoureusement la convergence de la solution du système $(\mathcal{P}^\lambda)$ vers la solution du système $(\mathcal{P}^0)$ quand $\lambda \to 0$, le point-clé est l'estimation d'entropie-dissipation suivante :

$$0 \leq \mathbb{E}(t) + \int_0^t \mathbb{I}(\tau)\, d\tau \leq \mathbb{E}(0), \tag{1.13}$$

où la dissipation est toujours définie par (1.11), mais où cette fois-ci l'entropie relative est définie par rapport aux conditions de Dirichlet N^D, P^D, Ψ^D, relevées sur tout le domaine :

$$\begin{aligned} \mathbb{E}(t) = & \int_\Omega \Big(H(N(t)) - H(N^D) - \log(N^D)(N(t) - N^D) \Big)\, dx \\ & + \int_\Omega \Big(H(P(t)) - H(P^D) - \log(P^D)(P(t) - P^D) \Big)\, dx \\ & + \frac{\lambda^2}{2} \int_\Omega \Big| \nabla(\Psi(t) - \Psi^D) \Big|^2\, dx. \end{aligned} \tag{1.14}$$

En supposant que $\mathbb{E}(0)$ est bornée indépendamment de λ, cette inégalité (1.13) fournit des bornes uniformes en λ sur $(N^\lambda, P^\lambda, \Psi^\lambda)$ solution du système $(\mathcal{P}^\lambda)$, qui permettent d'obtenir de la compacité d'une famille de solutions approchées et de passer à la limite $\lambda \to 0$ dans le système.

1.2 Discrétisation du modèle de dérive-diffusion

La première discrétisation du modèle de Van Roosbroeck (1.1)–(1.3) a été proposée en 1964 par H. Gummel [100] et améliorée quelques années plus tard par D. Scharfetter et H. Gummel [158]. Mentionnons également les travaux de A. M. Il'in [114] où le même type de flux que dans [158] a été introduit pour des schémas différences finies en une dimension d'espace. Le schéma de Scharfetter-Gummel a ensuite été interprété comme une méthode d'éléments finis mixtes et étendu au cas multidimensionnel par F. Brezzi, L. D. Marini et

P. Pietra [33, 34]. Plusieurs autres extensions au cas multidimensionnel ont par la suite été proposées ; pour plus de détails, on pourra se référer à l'article de revue de F. Brezzi et coll. [32].
La méthode d'éléments finis mixtes proposée pour le modèle de Van Roosbroeck a aussi été adaptée au cas d'une diffusion non linéaire par F. Arimburgo, C. Baiocchi, L. Marini dans [9] et par A. Jüngel dans [118] pour le problème unidimensionnel, et par A. Jüngel et P. Pietra dans [122] pour le modèle en dimension 2.
Plus récemment, des méthodes de volumes finis ont été proposées : ainsi, C. Chainais-Hillairet et Y. J. Peng étudient dans [53] un schéma volumes finis pour le système de dérive-diffusion en 1-D, qu'ils étendent au cas multidimensionnel dans [52, 54]. Citons enfin le schéma préservant l'asymptotique en temps long introduit par C. Chainais-Hillairet et F. Filbet dans [51].
Dans la partie I, nous considérons une discrétisation en temps de type Euler implicite et une discrétisation en espace de type volumes finis, qui utilise les flux de Scharfetter-Gummel. L'objectif de cette partie est de montrer que cette discrétisation particulière du système de dérive-diffusion standard en préserve les deux asymptotiques décrites ci-dessus : l'asymptotique en temps long et la limite quasi-neutre. Avant de présenter plus précisément les résultats obtenus, nous commençons par introduire le travail effectué en rappelant le principe des schémas préservant des asymptotiques.

1.3 Schémas préservant des asymptotiques

On considère un problème $(\mathcal{P}^\varepsilon)$ dépendant d'un paramètre $\varepsilon > 0$. On suppose que le régime asymptotique $\varepsilon \to 0$ dans $(\mathcal{P}^\varepsilon)$ est connu et conduit à un problème limite $(\mathcal{P}^0)$. On désigne par $(\mathcal{P}_\delta^\varepsilon)$ une discrétisation du problème $(\mathcal{P}^\varepsilon)$, où δ représente le paramètre de discrétisation.
On dit que le schéma numérique $(\mathcal{P}_\delta^\varepsilon)$ préserve l'asymptotique $\varepsilon \to 0$ s'il fournit une discrétisation stable du problème $(\mathcal{P}^\varepsilon)$ pour toute valeur de $\varepsilon > 0$ et si lorsque $\varepsilon \to 0$, à δ fixé, il conduit à un schéma $(\mathcal{P}_\delta^0)$ consistant avec le problème limite $(\mathcal{P}^0)$. Schématiquement, il s'agit de faire commuter le diagramme suivant :

$$\begin{array}{ccc} \mathcal{P}^\varepsilon & \xrightarrow{\quad\varepsilon \to 0\quad} & \mathcal{P}^0 \\ \delta \to 0 \uparrow & & \uparrow \delta \to 0 \\ \mathcal{P}_\delta^\varepsilon & \xrightarrow[\quad\varepsilon \to 0\quad]{} & \mathcal{P}_\delta^0 \end{array}$$

La notion de schéma « préservant l'asymptotique » (ou *asymptotic preserving*, abrégé par AP) a été introduite par S. Jin dans [117], dans le contexte des limites diffusives dans les systèmes cinétiques. L'idée est de construire des schémas numériques valides aussi bien dans le régime cinétique que dans le régime fluide, indépendamment du paramètre de discrétisation choisi.

Dans la partie I, nous nous intéressons à une discrétisation du système de dérive-diffusion isotherme pour les semi-conducteurs qui d'une part préserve l'asymptotique en temps long, ce qui revient à prendre $\varepsilon = 1/t$, et d'autre part préserve la limite quasi-neutre, ce qui revient à prendre $\varepsilon = \lambda$, la longueur de Debye adimensionnée. Un certain nombre de schémas AP pour la limite quasi-neutre ont été développés pour le système d'Euler-Poisson [65] et pour le système de Vlasov-Poisson [67, 68]. Concernant la préservation de l'asymptotique en temps long, H. Gajewski et K. Gärtner prouvent une inégalité d'énergie analogue à (1.12) pour une semi-discrétisation temporelle implicite, et montrent une propriété de dissipativité du schéma de Scharfetter-Gummel de type éléments finis [92]. Pour le système de dérive-diffusion non linéaire, C. Chainais-Hillairet et F. Filbet proposent une nouvelle discrétisation de type volumes finis en espace et semi-implicite en temps [51], et prouvent que le schéma dans le cas évolutif converge, quand le nombre de pas de temps tend vers l'infini, vers un schéma pour le modèle stationnaire.

1.4 Présentation des résultats

Dans la partie I, nous considérons donc une discrétisation du système de dérive-diffusion standard (1.1)–(1.3) par un schéma de type volumes finis implicite en temps avec une approximation de Scharfetter-Gummel des flux de convection-diffusion. Le schéma étant implicite en temps, il s'écrit comme un système d'équations non linéaires à résoudre à chaque pas de temps. Nous nous assurons donc de l'existence de solutions à ce système ; c'est l'objet du théorème 1.2.1 (chapitre 1). La preuve s'appuie sur l'étude d'une version semi-implicite en temps du schéma et sur une application du théorème de Brouwer. Notons que l'utilisation de ce théorème de point fixe ne permet pas de prouver l'unicité de la solution, qui reste à ce jour une question ouverte.

Comportement en temps long

Le but de la partie I est de prouver que le schéma de Scharfetter-Gummel tout implicite préserve des asymptotiques. Nous nous intéressons dans le chapitre 1 à l'asymptotique en temps long. Nous adaptons pour cela au niveau discret la méthode d'énergie–dissipation utilisée au niveau continu [92], en suivant la même démarche que dans [51]. Le point-clé de l'étude du comportement en temps long des solutions numériques est le résultat suivant, correspondant à la proposition 1.4.1 du chapitre 1. C'est un analogue discret de l'inégalité (1.12) :

Proposition 1. *En supposant que les conditions au bord de Dirichlet N^D, P^D et Ψ^D sont à l'équilibre thermique (c'est-à-dire satisfont la condition de compatibilité* (1.7)*), on a pour tout $n \geq 0$:*

$$0 \leq \mathcal{E}^{n+1} + \Delta t\, \mathcal{I}^{n+1} \leq \mathcal{E}^n, \tag{1.15}$$

où $\mathcal{E}^n$ et $\mathcal{I}^n$ sont respectivement des approximations de l'énergie relative $\mathcal{E}$ définie par (1.10) *et de sa dissipation $\mathcal{I}$ définie par* (1.11) *au temps $t^n = n\,\Delta t$.*

La preuve de cette proposition utilise fortement des propriétés techniques propres à la discrétisation de Scharfetter-Gummel des flux. Elle est également fondée sur le caractère totalement implicite en temps du schéma. Cependant, en utilisant une discrétisation semi-implicite analogue à celle considérée dans [51], nous pourrions encore obtenir une estimation d'énergie–dissipation, mais avec une contrainte sur le pas de temps Δt liée à la longueur de Debye adimensionnée λ.
En utilisant l'inégalité (1.15), nous pouvons alors prouver le théorème suivant (Theorem 1.1.1, résultat principal du chapitre 1), en suivant la même démonstration que dans [51, Theorem 2.2] :

Théorème 2. *Sous les hypothèses suivantes :*

1. *il existe m, $M > 0$ tels que $0 < m \leq N_0, P_0, N^D, P^D \leq M$,*
2. *les conditions de Dirichlet satisfont* (1.7)*,*
3. *le dopage C est nul sur Ω,*

la solution approchée $(N_\delta, P_\delta, \Psi_\delta)$ obtenue par le schéma de Scharfetter-Gummel tout implicite converge en temps long vers l'approximation $(N_\delta^{eq}, P_\delta^{eq}, \Psi_\delta^{eq})$ de l'équilibre thermique obtenue avec le schéma pour le modèle stationnaire proposé dans [51].

Nous remarquons que l'hypothèse 3 de dopage nul n'est pas nécessaire pour prouver l'inégalité d'énergie–dissipation discrète (1.15). Elle est cependant cruciale pour déduire de l'hypothèse 1 une borne inférieure uniforme sur N et P pour tout temps, qui est indispensable dans notre preuve. À la fin du chapitre 1, nous présentons quelques simulations numériques qui montrent bien la convergence des solutions approchées vers une approximation de l'équilibre thermique, et ce même si le dopage C est non nul. Nous comparons les résultats avec ceux obtenus en utilisant une autre discrétisation en espace, qui ne préserve pas l'équilibre thermique et apparaît beaucoup moins efficace pour refléter correctement le comportement en temps long des solutions.

Limite quasi-neutre

Dans le chapitre 2, nous nous intéressons à la stabilité à la limite quasi-neutre du schéma de Scharfetter-Gummel implicite. Le but est de prouver que ce schéma converge vers une solution du système (1.1)–(1.3) pour toute valeur de la longueur de Debye adimensionnée. Le résultat principal est donné par le théorème suivant (théorème 2.1.2 du chapitre 2) :

Théorème 3. *On suppose que :*

1. *N_0, $P_0 \in L^\infty(\Omega)$, N^D, $P^D \in L^\infty(\Omega) \cap H^1(\Omega)$, $\Psi^D \in H^1(\Omega)$,*
2. *il existe m, $M > 0$ tels que $0 < m \leq N_0, P_0, N^D, P^D \leq M$,*
3. *le dopage C est nul sur Ω,*
4. *les conditions au bord vérifient la quasi-neutralité : $N^D = P^D$,*
5. *les conditions initiales vérifient la quasi-neutralité : $N_0 = P_0$.*

Alors pour toute valeur de $\lambda > 0$, il existe des fonctions N, $P \in L^2(0,T;H^1(\Omega))$ et $\Psi \in \left(L^2(0,T;H^1(\Omega))\right)^d$ tels que la solution $(N_\delta, P_\delta, \Psi_\delta)$ obtenue avec le schéma de Scharfetter-Gummel implicite satisfasse quand $\delta \to 0$, à l'extraction de sous-suites près,

$$N_\delta \to N, \quad P_\delta \to P \text{ dans } L^2(0,T;H^1(\Omega)) \text{ fortement,}$$
$$\Psi_\delta \rightharpoonup \Psi \text{ dans } \left(L^2(\Omega\times(0,T))\right)^d \text{ faiblement,}$$
$$dN_\delta \rightharpoonup \nabla N, \quad dP_\delta \rightharpoonup \nabla P, \quad d\Psi_\delta \rightharpoonup \nabla\Psi \text{ dans } \left(L^2(\Omega\times(0,T))\right)^d \text{ faiblement,}$$

où dN_δ, dP_δ et $d\Psi_\delta$ sont des approximations des gradients de N, P et Ψ obtenues avec le schéma.
De plus, la limite (N,P,Ψ) est une solution faible du système (1.1)–(1.3).

Pour prouver ce résultat, nous suivons une démarche similaire à celle adoptée par C. Chainais-Hillairet, J. G. Liu et Y. J. Peng dans [52, 54] : nous commençons par montrer un certain nombre d'estimations a priori discrètes – estimations $L^\infty(\Omega\times(0,T))$, BV faible, $L^2(0,T;H^1(\Omega))$ sur les densités N_δ et P_δ, et estimation $L^2(0,T;H^1(\Omega))$ sur le potentiel Ψ_δ – qui permettent d'obtenir la compacité d'une suite de solutions approchées en utilisant des arguments classiques. Toutefois, les schémas étudiés dans [52, 54] sont uniquement considérés dans le cas particulier où $\lambda = 1$. Puisque nous souhaitons ici obtenir la convergence pour toutes valeurs de $\lambda > 0$ et analyser la limite quasi-neutre $\lambda \to 0$, il est crucial que toutes les estimations a priori discrètes soient indépendantes de λ. Nous nous contentons ici d'étudier la convergence pour tout $\lambda > 0$, l'étude complète du passage à la limite quasi-neutre au niveau discret sera l'objet d'un travail futur.
Nous ne pouvons pas appliquer directement les mêmes stratégies que dans [52, 54]. La clé pour obtenir ces estimations a priori uniformes en λ est d'adapter au cadre discret la méthode d'entropie utilisée dans le contexte continu [94, 95, 121]. La proposition suivante (proposition 2.2.1, chapitre 2) nous fournit un analogue discret de l'estimation (1.13) :

Proposition 4. *On suppose que la solution du schéma numérique vérifie l'estimation L^∞ suivante :*

$$0 < m \le N_\delta, P_\delta \le M. \tag{1.16}$$

Alors il existe une constante C_E dépendant uniquement des données Ω, T, N^D, P^D, Ψ^D, N_0, P_0, et des bornes m et M, telle que pour tout $\lambda > 0$,

$$\frac{\mathbb{E}^{n+1} - \mathbb{E}^n}{\Delta t} + \frac{1}{2}\,\mathbb{I}^{n+1} \le C_E,$$

où $\mathbb{E}^n$ et $\mathbb{I}^n$ sont des versions discrètes respectivement de l'entropie relative $\mathbb{E}$ définie par (1.14) *et de sa dissipation définie par* (1.11).
Si de plus les conditions initiales vérifient la quasi-neutralité ($N_0 = P_0$), alors

$$\sum_{n=0}^{N_T-1} \Delta t\, \mathbb{I}^{n+1} \le C_E(1+\lambda^2).$$

La preuve de ce résultat s'appuie à nouveau fortement sur les propriétés particulières des flux de Scharfetter-Gummel. De plus, le choix d'une discrétisation totalement implicite en temps est cette fois-ci cruciale pour obtenir des estimations indépendantes de λ. En effet, si nous choisissons une discrétisation semi-implicite en temps comme c'est le cas dans [51, 52, 54], cela induit une condition de stabilité de la forme $\Delta t \leq C\,\lambda^2$, et nous ne pouvons donc pas utiliser une telle discrétisation pour des petites valeurs de λ. Remarquons enfin que nous ne supposons pas le dopage nul dans cette proposition 4. Cependant, pour montrer l'inégalité d'entropie, nous avons besoin de l'estimation L^∞ (1.16) qui est garantie dans le cas d'un dopage C nul.
Une fois la convergence de $(N_\delta, P_\delta, \Psi_\delta)$ vers (N, P, Ψ) établie, il reste à passer à la limite dans le schéma pour vérifier que la limite (N, P, Ψ) obtenue est bien solution faible du système (1.1)–(1.3). Dans ce cadre, la principale difficulté provient du fait que les parties convective et diffusive sont discrétisées ensemble dans le flux de Scharfetter-Gummel. Il faut donc réussir à récupérer à la limite chacune de ces deux parties séparément.

Le résultat principal du chapitre 2 constitue la première étape pour montrer que le schéma de Scharfetter-Gummel préserve l'asymptotique quasi-neutre : dans le diagramme présenté page 7, nous avons montré la convergence $\mathcal{P}_\delta^\lambda \longrightarrow \mathcal{P}^\lambda$ quand $\delta \to 0$, pour toute valeur de $\lambda > 0$. L'étude du caractère AP du schéma reste à compléter, comme expliqué dans la sous-section suivante ; nous nous contentons de donner quelques résultats numériques à la fin du chapitre 2 qui semblent bien mettre en lumière la convergence du schéma indépendamment de λ.

2 Schémas préservant l'asymptotique en temps long pour des équations paraboliques non linéaires

Dans cette section, nous introduisons les travaux présentés dans la partie II. Lors de notre étude du schéma de Scharfetter-Gummel pour le système de dérive-diffusion, nous avons constaté en comparant les résultats numériques obtenus avec d'autres schémas volumes finis que ce schéma était très efficace, tant du point de vue de la préservation de l'asymptotique en temps long que de la précision. Cependant, le flux de Scharfetter-Gummel n'est défini que dans le cas d'une diffusion linéaire. Dans la partie II, l'idée est donc de proposer une extension de la définition du schéma de Scharfetter-Gummel, et plus généralement un schéma préservant le comportement en temps long, pour une équation de convection-diffusion non linéaire générale de la forme

$$\partial_t u = \operatorname{div}(f(u)\,\nabla V(x) + \nabla r(u)), \quad x \in \Omega \subset \mathbb{R}^d, \quad t > 0, \tag{2.1}$$

où $u : \mathbb{R}^+ \times \Omega \to \mathbb{R}^+$ est l'inconnue, V est un potentiel donné, r et f sont des fonctions régulières données. Notons que des généralisations du schéma de Scharfetter-Gummel au cas non linéaire ont déjà été introduites par R. Eymard, J. Fuhrmann et K. Gärtner [83], ainsi que par A. Jüngel et P. Pietra [118, 122], mais ces travaux ne concernent pas la préservation de l'asymptotique en temps long.

Dans cette section, nous précisons d'abord le cadre général considéré, en rappelant notamment les résultats et les méthodes utilisées au niveau continu pour étudier le comportement en temps long d'équations de convection-diffusion non linéaires de la forme (2.1). Nous donnons également des exemples de problèmes physiques décrits par de telles équations. Dans un deuxième temps, nous proposons un état de l'art non exhaustif des nombreuses méthodes numériques proposées pour approcher des équations de ce type et nous motivons nos travaux dans ce contexte. Enfin, nous présentons les résultats obtenus : dans le chapitre 3, nous proposons une extension du schéma de Scharfetter-Gummel pour une équation de convection-diffusion avec diffusion non linéaire et convection linéaire, et nous étudions sa convergence dans le cas non dégénéré. Constatant une dégradation des résultats numériques obtenus lorsque la diffusion dégénère, nous proposons dans le chapitre 4 une nouvelle discrétisation de l'équation (2.1), incluant également le cas d'une convection non linéaire.

2.1 Présentation du cadre général

Dans la partie II, nous nous intéressons à des schémas volumes finis préservant le comportement en temps long. Nous nous plaçons pour cette étude dans le cadre général d'un problème parabolique non linéaire, éventuellement dégénéré :

$$\begin{cases} \partial_t u = \operatorname{div}(f(u)\nabla V(x) + \nabla r(u)), \quad x \in \Omega, \quad t > 0, \\ \\ u(t=0,x) = u_0(x), \end{cases} \tag{2.2}$$

où $\Omega \subset \mathbb{R}^d$ est un domaine ouvert borné ou l'espace $\mathbb{R}^d$ tout entier. Si Ω est un domaine borné, nous imposons de plus des conditions aux limites qui dépendent du problème particulier considéré (nous considèrerons essentiellement des conditions de Neumann homogènes ou de Dirichlet). Dans le problème (2.2), $u \geq 0$ est une densité, f est une fonction donnée et $r \in \mathcal{C}^1(\mathbb{R}^+)$ est telle que $r'(u) \geq 0$ et $r'(u)$ peut s'annuler pour certaines valeurs de u (le problème (2.2) est alors dit dégénéré). De plus, nous supposons que r et f sont telles que $f(u) \geq 0$ et qu'il existe une fonction h telle que

$$r'(s) = h'(s)\, f(s). \tag{2.3}$$

Cette hypothèse signifie que le problème considéré a une structure correspondant à une énergie ou une entropie, ou plus généralement à une fonctionnelle de Lyapunov. En effet, l'équation (2.2) peut se réécrire sous la forme

$$\partial_t u = \operatorname{div}\left(f(u)\nabla\left(V(x) + h(u)\right)\right) \tag{2.4}$$

et en multipliant cette équation par $(V + h(u))$ et en intégrant sur Ω, on obtient si on considère des conditions aux limites de Neumann homogènes :

$$\frac{dE(t)}{dt} = -\mathcal{I}(t) \leq 0, \tag{2.5}$$

où la fonctionnelle E est donnée par

$$E(t) = \int_\Omega (V\,u + H(u))\,dx,$$

H étant une primitive de h, et la dissipation $\mathcal{I}$ par

$$\mathcal{I}(t) = \int_\Omega f(u)\,|\nabla(V + h(u))|^2\,dx.$$

Cette structure particulière permet d'étudier le comportement en temps long du problème (2.2), en utilisant des méthodes d'entropie–dissipation d'entropie dont le principe général a été expliqué dans la section précédente. Nous commençons par repréciser les principaux points de cette étude au niveau continu, dans le cas d'une convection linéaire ($f(u) = u$). Nous verrons ensuite les résultats obtenus dans le cas d'une diffusion linéaire ($r(u) = u$) et d'une convection non linéaire.

Cas d'une convection linéaire

Nous commençons par rappeler ici des résultats démontrés par J. A. Carrillo, A. Jüngel, P. A. Markowich, G. Toscani et A. Unterreiter dans [45]. Cet article concerne le comportement en temps long du problème de convection-diffusion général suivant, qui correspond à (2.2) dans le cas d'une convection linéaire :

$$\partial_t u = \mathrm{div}(u\nabla V(x) + \nabla r(u)), \quad x \in \Omega, \quad t > 0, \tag{2.6}$$

avec pour condition initiale $u_0 \in L^1_+(\Omega) := \{u \in L^1(\Omega);\, u \geq 0\}$ de masse $M = \int_\Omega u_0(x)\,dx$. Cette équation est complétée par une condition de décroissance quand $|x| \to \infty$ si Ω est l'espace $\mathbb{R}^d$ tout entier ou par une condition de flux nul sur la frontière $\partial\Omega$ si Ω est borné. La fonction $r : \mathbb{R}_+ \to \mathbb{R}$ appartient à $\mathcal{C}^2(\mathbb{R}_+)$, est croissante et vérifie $r(0) = 0$. Dans ce cas, la fonction h est définie de la manière suivante :

$$h(s) = \int_1^s \frac{r'(\tau)}{\tau}\,d\tau, \quad s \in]0, \infty[, \tag{2.7}$$

et est supposée appartenir à $L^1_{loc}([0, +\infty))$, de telle sorte que la fonction H donnée par

$$H(s) = \int_0^s h(\tau)\,d\tau, \quad s \in [0, +\infty[\tag{2.8}$$

est bien définie et vérifie $H'(s) = h(s)$ pour tout $s \geq 0$.
Comme expliqué précédemment, la première étape consiste à étudier les solutions stationnaires de (2.6). Elles vérifient :

$$u^{eq}\,\nabla V(x) + \nabla r(u^{eq}) = 0 \quad \text{et} \quad \int_\Omega u^{eq}(x)\,dx = M,$$

ce qui peut se réécrire en utilisant la définition (2.7) de h :

$$u^{eq}\,\nabla\,(V(x) + h(u^{eq})) = 0 \quad \text{et} \quad \int_\Omega u^{eq}(x)\,dx = M.$$

Si $u^{eq} > 0$ sur Ω, alors on obtient

$$V(x) + h(u^{eq}(x)) = C \quad \forall x \in \Omega,$$

où $C \in \mathbb{R}$ est telle que $\int_\Omega u^{eq}(x)\,dx = M$.
De manière plus rigoureuse, en considérant la fonctionnelle d'entropie suivante :

$$E(u) = \int_\Omega \left(V(x)\,u + H(u)\right)\,dx,$$

une fonction $u^{eq,M} \in L^1(\Omega)$ est une solution à l'équilibre de (2.6) si et seulement si c'est un minimiseur de E dans

$$\mathcal{C} = \left\{ u \in L^1(\Omega),\ \int_\Omega u(x)\,dx = M \right\}.$$

Sous des hypothèses de régularité sur le potentiel V, l'existence et l'unicité de l'équilibre sont prouvées dans [45].
Ensuite, en utilisant la méthode d'entropie–dissipation décrite précédemment, la décroissance exponentielle de l'entropie relative

$$\mathcal{E}(t) = E(u(t)) - E(u^{eq,M})$$

est prouvée, en utilisant la décroissance exponentielle de la dissipation d'entropie

$$\mathcal{I}(t) = -\frac{d\mathcal{E}(t)}{dt} = \int_\Omega u(x,t)\ \left|\nabla\left(V(x) + h(u(x,t))\right)\right|^2\,dx.$$

Finalement, grâce à une inégalité de Csiszar-Kullback généralisée, la convergence en norme L^1 de la solution $u(t,x)$ de (2.6) vers l'équilibre $u^{eq,M}(x)$ à un taux exponentiel quand $t \to \infty$ est démontrée.
Nous présentons à présent deux exemples d'équations issues de la physique qui s'inscrivent dans ce cadre général et auxquels nous nous intéressons dans la partie II.

Exemple (Équation des milieux poreux)**.** L'écoulement d'un gaz parfait dans un milieu poreux homogène peut être décrit par le modèle de Darcy-Leibenzon-Muskat suivant :

$$\begin{cases} \partial_t v = \Delta v^m & \text{sur } \mathbb{R}^d \times (0,T), \\ v(x,0) = v_0(x) & \text{sur } \mathbb{R}^d, \end{cases} \tag{2.9}$$

où la fonction v représente la densité de gaz dans le milieu poreux et $m > 1$ est une constante physique.
En faisant un changement de variables dépendant du temps, le problème peut se réécrire sous la forme de l'équation de Fokker-Planck non linéaire suivante :

$$\begin{cases} \partial_t u = \operatorname{div}(xu + \nabla u^m) & \text{sur } \mathbb{R}^d \times (0,T), \\ u(x,0) = u_0(x) & \text{sur } \mathbb{R}^d. \end{cases} \tag{2.10}$$

Plus précisément, si v est solution de (2.9), alors

$$u(x,t) = e^{dt}\, v\left(k\,(e^{t/k}-1), e^t x\right),$$

où $k = (d(m-1)+2)^{-1}$, est solution de (2.10), et réciproquement si u est solution de (2.10), alors

$$v(x,t) = \left(1+\frac{t}{k}\right)^{-dk} u\left(k\,\log\left(1+\frac{t}{k}\right), \left(1+\frac{t}{k}\right)^{-k}\right)$$

est solution de (2.9).
J. A. Carrillo et G. Toscani étudient dans [48] le comportement en temps long de l'équation des milieux poreux en utilisant la reformulation (2.10), qui correspond à la forme générale (2.6) avec $V(x) = |x|^2/2$ et $r(u) = u^m$. Ils prouvent que l'unique solution stationnaire de (2.10) est la distribution de Barenblatt-Pattle suivante :

$$u^{eq}(x) = \left(C - \frac{m-1}{2m}|x|^2\right)_+^{1/(m-1)},$$

où C est une constante telle que u^{eq} ait la même masse que la donnée initiale u_0. En utilisant la méthode d'entropie–dissipation décrite ci-dessus, la convergence de $u(x,t)$ vers $u^{eq}(x)$ à vitesse exponentielle quand $t \to \infty$ est démontrée dans [48].

Exemple (Système de dérive-diffusion pour les semi-conducteurs)**.** Nous avons déjà présenté dans la section 1 le modèle de dérive-diffusion linéaire pour les semi-conducteurs (1.1). Il est supposé dans ce modèle que les porteurs de charges se comportent comme un gaz parfait : la pression de trous et d'électrons est donnée par $r(s) = sT(s)$, où T est la température, supposée la même pour les électrons et pour les trous. Dans le modèle standard (1.1), la température est supposée constante (modèle isotherme) et donc $r(s) = s$. Si maintenant la dépendance de T en les densités de porteurs de charge est prise en compte, nous pouvons considérer que la pression est donnée par $r(s) = s^m$, avec $m > 1$ ($m = 5/3$ correspond par exemple au modèle isentropique). On obtient alors le système de dérive-diffusion non linéaire suivant :

$$\left\{\begin{array}{l} \partial_t N - \mathrm{div}(\nabla r(N) - N\nabla\Psi) = 0 \\ \partial_t P - \mathrm{div}(\nabla r(P) + P\nabla\Psi) = 0 \quad \text{dans } \Omega\times(0,T), \\ \lambda^2\Delta\Psi = N - P - C \end{array}\right. \tag{2.11}$$

où les notations considérées sont les mêmes que dans la section 1. Ce système est complété par des conditions initiales $N_0(x)$ et $P_0(x)$ (1.3) et par des conditions aux limites mixtes (1.2) : conditions de Dirichlet N^D, P^D, Ψ^D au niveau des contacts ohmiques Γ^D et conditions de Neumann homogènes sur les parties isolées Γ^N. Les deux équations de continuité sur les densités N et P correspondent à la forme générale (2.6) avec $r(s) = s^m$ la fonction de pression.
Le système stationnaire est étudié dans [138]. Il admet une solution $(N^{eq}, P^{eq}, \Psi^{eq})$, qui est unique si de plus :
(2.12)

$$h(N^{eq}) - \Psi^{eq} \left\{\begin{array}{lll} = & \alpha_N & \text{si } N^{eq} > 0, \\ \geq & \alpha_N & \text{si } N^{eq} = 0 \end{array}\right. \quad \text{et} \quad h(P^{eq}) + \Psi^{eq} \left\{\begin{array}{lll} = & \alpha_P & \text{si } P^{eq} > 0, \\ \geq & \alpha_P & \text{si } P^{eq} = 0 \end{array}\right.$$

et si les conditions de Dirichlet satisfont (2.12) et la condition de compatibilité (si N^D, $P^D > 0$) :

$$h(N^D) + h(P^D) = \alpha_N + \alpha_P.$$

Dans ce cas, l'équilibre thermique est défini par :

$$\left\{ \begin{array}{l} \lambda^2 \Delta \Psi^{eq} = g(\alpha_N + \Psi^{eq}) - g(\alpha_P - \Psi^{eq}) - C \\ N^{eq} = g(\alpha_N + \Psi^{eq}), \quad P^{eq} = g(\alpha_P - \Psi^{eq}) \end{array} \right. \quad \text{sur } \Omega, \tag{2.13}$$

où g est l'inverse généralisé de h :

$$g(s) = \left\{ \begin{array}{lll} h^{-1}(s) & \text{si} & h(0_+) < s < +\infty, \\ 0 & \text{si} & s \leq h(0_+). \end{array} \right.$$

Le comportement en temps long de ce système est étudié par A. Jüngel dans [119] en utilisant la fonctionnelle d'énergie relative

$$\begin{aligned} \mathcal{E}(t) = \int_\Omega \Big(& H(N(t)) - H(N^{eq}) - h(N^{eq})(N(t) - N^{eq}) \\ & + H(P(t)) - H(P^{eq}) - h(P^{eq})(P(t) - P^{eq}) \\ & + \frac{\lambda^2}{2} \left| \nabla(\Psi(t) - \Psi^{eq}) \right|^2 \Big) dx \end{aligned}$$

et la dissipation d'énergie

$$\mathcal{I}(t) = \int_\Omega \Big(N(t) \left| \nabla(h(N(t)) - \Psi(t)) \right|^2 + P(t) \left| \nabla(h(P(t)) + \Psi(t)) \right|^2 \Big) dx,$$

où les fonctions h et H sont définies par (2.7) et (2.8) respectivement.
Dans ce cas non linéaire, le taux de convergence vers l'équilibre n'est pas calculé, contrairement au cas isotherme.

Cas d'une convection non linéaire

Nous nous intéressons maintenant au cas où le terme de convection est non linéaire. À notre connaissance, il n'existe pas d'étude générale de tels modèles. Nous présentons donc des résultats obtenus dans le cas particulier d'un modèle proposé par G. Kaniadakis et P. Quarati [124, 125] pour décrire des gaz de bosons ou de fermions. Dans ce modèle, une correction est apportée au terme de convection de l'équation de Fokker-Planck classique pour décrire des particules interagissant avec un principe d'exclusion, comme c'est le cas pour les fermions ou les bosons. Cette équation correspond à (2.2) avec une diffusion linéaire ($r(u) = u$) et une convection non linéaire ($f(u) = u(1 + ku)$, $V(x) = |x|^2/2$) :

$$\partial_t u = \operatorname{div}(xu(1 + ku) + \nabla u), \quad x \in \mathbb{R}^d, \quad t > 0, \tag{2.14}$$

avec $k = 1$ dans le cas des bosons et $k = -1$ dans le cas des fermions. Le comportement en temps long de ce modèle a été étudié dans les deux cas en 1D [47], dans toutes les

dimensions pour les fermions [46] et dans le cas 3D pour les bosons [160]. La solution stationnaire de (2.14) est donnée par les distributions de Fermi-Dirac ($k = -1$) et Bose-Einstein ($k = 1$) :

$$u^{eq}(x) = \frac{1}{\beta\, e^{|x|^2/2} - k},$$

où $\beta \geq 0$ est telle que u^{eq} a la même masse que u_0. La fonctionnelle d'entropie est définie par

$$E(u) = \int_{\mathbb{R}^d} \left(\frac{|x|^2}{2}\, u + u \log(u) - k(1+ku)\log(1+ku) \right) dx,$$

et la dissipation d'entropie correspondante est donnée par

$$\mathcal{I}(t) = \int_{\mathbb{R}^d} u\,(1+ku) \left| \nabla \left(\frac{|x|^2}{2} + \log\left(\frac{u}{1+ku} \right) \right) \right|^2 dx.$$

En une dimension d'espace pour les bosons et en toute dimension pour les fermions, la convergence à un taux exponentiel de $u(x,t)$ vers $u^{eq}(x)$ quand $t \to \infty$ est démontrée par J. A. Carrillo, P. Laurençot, J. Rosado, F. Salvarani [47, 46] en utilisant la méthode d'entropie–dissipation.

Concernant le modèle 3D pour les bosons, G. Toscani [160] met en évidence l'existence d'une masse critique M^* dans ce cas. Pour des densités initiales dont la masse $M \ll M^*$ est sous-critique, la norme L^2 de la solution reste bornée uniformément en temps, alors que pour des masses sur-critiques $M \gg M^*$, la solution se concentre et explose en temps fini (formation d'un condensat de Bose-Einstein). Ces résultats sont fondés sur des inégalités de type Nash pour contrôler l'évolution de la norme L^2 de la solution.

Plus récemment, N. Ben Abdallah, I. Gamba et G. Toscani [15] ont étudié une classe d'équations de type Fokker-Planck plus générales :

$$\partial_t u = \operatorname{div}(x\, f(u) + \nabla u) \quad \text{sur } \mathbb{R}^d \times (0,T), \tag{2.15}$$

où la fonction f est strictement croissante, et vérifie $f(0) = 0$, et

$$\int_1^{+\infty} \frac{ds}{f(s)} \leq K < \infty, \quad \lim_{s \to 0^+} \frac{f(s)}{s} = 1,$$

par exemple $f(s) = s\,(1 + s^N)$, $N > 0$. La fonction h vérifiant (2.3) est alors définie par

$$h(s) = -\int_s^{+\infty} \frac{d\tau}{f(\tau)}.$$

Les états stationnaires de (2.15) sont donnés par

$$u^{eq,M}(x) = h^{-1}\left(-\frac{|x|^2}{2} - \lambda \right), \tag{2.16}$$

où $\lambda \geq 0$ est telle que $\int_{\mathbb{R}^d} u^{eq,M}(x)\,dx = M$. Dans [15], les auteurs s'intéressent au problème de minimisation de la fonctionnelle d'entropie

$$E(u) = \int_{\mathbb{R}^d} \left(\frac{|x|^2}{2}\, u + H(u) \right) dx$$

dans l'ensemble

$$\mathcal{C}_M = \left\{ u \in \mathcal{M}_b^+(\mathbb{R}^d);\ \int_{\mathbb{R}^d} u(x)\,dx = M \right\}.$$

Ils prouvent l'existence d'une masse critique M^* telle que si la masse initiale M est plus petite que M^*, alors l'unique minimiseur de E dans $\mathcal{C}_M$ est $u^{eq,M}$ défini par (2.16), mais si par contre $M > M^*$, alors l'unique minimiseur de E dans $\mathcal{C}_M$ a une partie singulière localisée en 0.

2.2 Discrétisation du problème : état de l'art et motivations

Il existe une littérature significative concernant les schémas numériques pour approcher les solutions d'équations de convection-diffusion non linéaires. En effet, ces problèmes interviennent dans de nombreux processus en science et technologie : simulation de réservoirs de pétrole, hydrogéologie, protection environnementale (simulation de la propagation de polluants),... Dès lors, nous commençons par dresser un bref panorama des méthodes existantes. Il existe d'abord de nombreux travaux relatifs aux méthodes d'éléments finis, appliquées notamment à des problèmes issus de l'ingénierie pétrolière : éléments finis linéaires par morceaux [13, 77, 115, 145, 146], éléments finis mixtes [7, 66]. Les schémas volumes finis aussi s'avèrent efficaces dans le cas d'équations paraboliques dégénérées, qu'il s'agisse de méthodes *cell-centered* [85, 87] ou *vertex-centered* [147], qui permettent d'obtenir une bonne approximation des gradients. Des méthodes de Galerkin discontinues [58, 166] ainsi que des schémas combinés volumes finis–éléments finis ont également été analysés [6, 88]. Nous mentionnons par ailleurs des discrétisations de type différences finies proposées dans [81, 82]. Plus récemment, des schémas cinétiques reposant sur une analogie avec des modèles de BGK (Bhatnagar, Gross et Krook) discrets ont été introduits [8, 26], ainsi que des discrétisations fondées sur la méthode des caractéristiques [56, 123]. D'autres approches se fondent sur des méthodes de splitting permettant un traitement séparé des parties convectives et diffusives [39, 80] et sur l'utilisation du principe du maximum, en considérant une perturbation locale des données pour que le problème traité ne soit plus dégénéré [152]. Enfin, des méthodes d'ordre élevé ont été proposées : schémas MUSCL [130] ou ENO–WENO [49, 135].

Dans la partie II de ce manuscrit, notre but est de construire des schémas volumes finis qui préservent les états stationnaires, afin d'obtenir un comportement en temps long satisfaisant pour les solutions approchées. En effet, il apparaît que des schémas numériques fondés sur la préservation des états stationnaires fournissent un comportement très précis de la solution approchée en temps long. À notre connaissance, assez peu de travaux sont consacrés à l'étude de l'asymptotique en temps long des solutions numériques pour ce type de problème. Dans l'article [11] datant de 2003, A. Arnold et A. Unterreiter s'intéressent

à une semi-discrétisation en temps entièrement implicite pour des équations de Fokker-Planck linéaires et prouvent la décroissance exponentielle de l'entropie relative vers 0, ce qui leur permet de conclure à la convergence des solutions numériques vers l'unique état stationnaire. Pour leurs simulations numériques, ils utilisent une discrétisation en espace de type différences finies et mettent en évidence un phénomène de saturation de l'entropie relative au bout d'un certain temps, imputé à cette discrétisation spatiale. L. Gosse et G. Toscani [99] proposent un schéma explicite pour des équations de filtration 1D (équation des milieux poreux et équation de diffusion rapide) fondé sur une formulation alternative du problème utilisant le pseudo-inverse de la fonction de répartition de la densité. Leurs résultats numériques illustrent la capacité de cette méthode à restituer la convergence en temps long vers les solutions autosimilaires. C. Chainais-Hillairet et F. Filbet [51] s'intéressent quant à eux à une discrétisation volumes finis en espace, semi-implicite en temps, pour le système de dérive-diffusion non linéaire multidimensionnel (2.11). Ils prouvent la convergence de la solution numérique obtenue vers une approximation de l'équilibre thermique (2.13) quand $t \to +\infty$, en adaptant au niveau discret la méthode d'énergie–dissipation. Dans [43], J. A. Carrillo, M. Di Francesco et M. Gualdani prouvent la décroissance de l'entropie pour une semi-discrétisation implicite en temps d'équations de la forme (2.6). Enfin, M. Burger, J. A. Carrillo et M.-T. Wolfram [36] étudient une discrétisation en temps pour une classe d'équations de diffusion non linéaires fondée sur une reformulation du problème de diffusion comme un problème de transport optimal, et sur sa linéarisation. Ils prouvent que leur schéma semi-implicite préserve la décroissance exponentielle de la fonctionnelle d'énergie relative dans le cas d'équations de Fokker-Planck non linéaires avec potentiel uniformément convexe et appliquent leur méthode à l'équation des milieux poreux, à l'équation de diffusion rapide et au système de Patlak-Keller-Segel.
Dans la plupart de ces travaux, les résultats relatifs au comportement en temps long sont démontrés rigoureusement uniquement pour des semi-discrétisations temporelles [11, 36, 43]. De plus comme mentionné plus haut, lors de la discrétisation en espace du problème, un phénomène de saturation de l'entropie peut apparaître. Ce dernier élément souligne l'importance de considérer des discrétisations spatiales préservant les états stationnaires et la dissipation de l'entropie. Ce point de vue a été adopté dans [51]. Néanmoins, il apparaît que le schéma proposé perd en efficacité lorsque l'équation dégénère. Ainsi dans la partie II, nous proposons deux nouvelles discrétisations en espace de type volumes finis, construites de manière à préserver les états stationnaires et la dissipation de l'entropie. En outre, nous essayons également de construire une méthode qui reste précise même dans le cas dégénéré, afin de conserver les bonnes propriétés en temps long des solutions numériques y compris dans ce régime-là.

2.3 Présentation des résultats

Généralisation du schéma de Scharfetter-Gummel

En comparant les résultats numériques obtenus avec différents schémas volumes finis, il semble crucial que le flux numérique préserve les états d'équilibre pour obtenir un comportement en temps long de la solution approchée cohérent. Nous avons en particulier

constaté dans le chapitre 1 que le schéma de Scharfetter-Gummel est très efficace, aussi bien du point de vue de la préservation de l'asymptotique en temps long que de la précision (il est d'ordre deux en espace [132]). Cependant, il n'est défini que dans le cas d'une diffusion linéaire ($r(u) = u$ dans (2.6)). Par conséquent, l'idée du chapitre 3 est de construire un schéma généralisant le schéma de Scharfetter-Gummel dans le cas d'une diffusion non linéaire, qui peut éventuellement dégénérer, en s'attachant à conserver les états stationnaires afin de préserver l'asymptotique en temps long des solutions. Plusieurs extensions du schéma de Scharfetter-Gummel ont déjà été proposées. Ainsi, R. Eymard, J. Fuhrmann et K. Gärtner introduisent dans [83] un schéma valable dans le cas où la convection et la diffusion sont non linéaires, mais leur méthode conduit à résoudre un problème elliptique non linéaire à chaque interface, ce qui s'avère coûteux numériquement. A. Jüngel et P. Pietra ont aussi proposé un schéma pour le modèle de dérive-diffusion (2.11) dans [118, 122]. Leur définition est très proche de celle que nous proposons mais elle ne permet par contre pas de préserver les équilibres.
On considère dans le chapitre 3 le problème suivant :

$$\text{(2.17)} \qquad \begin{cases} \partial_t u = \operatorname{div}(\nabla r(u) - \mathbf{q}\, u), & (x,t) \in \Omega \times (0,T), \\ u(x,0) = u_0(x), & x \in \Omega, \end{cases}$$

complété avec des conditions au bord de Dirichlet : $u = \overline{u}$ sur $\partial\Omega \times (0,T)$. Notons que cette équation correspond à (2.6) avec $\mathbf{q} = \nabla V$. Dans un premier temps, nous construisons une extension du flux de Scharfetter-Gummel pour cette équation. Dans le cas linéaire $r(u) = u$, le flux de Scharfetter-Gummel classique permet d'approcher le flux

$$-\int_\sigma (\nabla u - \mathbf{q}\, u) \cdot \mathbf{n}_\sigma$$

à l'interface σ entre deux cellules du maillage du domaine Ω considéré ($\mathbf{n}_\sigma$ désigne la normale à l'interface σ). Dans un premier temps, nous définissons une extension de ce flux numérique dans le cas d'une diffusion linéaire avec un coefficient de viscosité $\varepsilon > 0$:

$$-\int_\sigma (\varepsilon \nabla u - \mathbf{q}\, u) \cdot \mathbf{n}_\sigma.$$

Finalement, nous construisons l'extension du schéma de Scharfetter-Gummel pour l'équation (2.17) en réécrivant $\nabla r(u)$ comme $r'(u)\nabla u$ et en considérant $r'(u)$ comme un coefficient de viscosité. La dernière étape consiste à définir une approximation de $r'(u)$ à l'interface σ, et c'est avec cette définition que nous assurons la préservation des états stationnaires par le flux. Le schéma que nous proposons reste valable dans le cas où le terme de diffusion s'annule, et dégénère en un schéma *upwind* classique (qui est donc d'ordre un en espace). Concernant la discrétisation en temps, nous considérons un schéma semi-implicite. Un tel choix nous permet d'obtenir un schéma inconditionnellement L^∞ stable (Proposition 3.3.1) et facile à mettre en oeuvre, puisqu'il conduit seulement à résoudre un système linéaire d'équations à chaque pas de temps.
Le résultat principal du chapitre 3 est résumé dans le théorème ci-après (qui correspond aux résultats de la proposition 3.4.1 et du théorème 3.4.1) :

Théorème 5. *On suppose que :*

1. $\overline{u} \in H^1(\Omega \times (0,T)) \cap L^\infty(\Omega \times (0,T))$, $u_0 \in L^\infty(\Omega)$,
2. *il existe* m, $M > 0$ *tels que* $0 < m \leq u_0$, $\overline{u} \leq M$,
3. $r \in \mathcal{C}^2(\mathbb{R})$ *est strictement croissante sur* $(0,+\infty)$, $r(0) = r'(0) = 0$, *et* $r'(s) \geq c_0\, s^{m-1}$, $m > 1$,
4. $\mathbf{q} \in \mathcal{C}^1(\overline{\Omega}, \mathbb{R}^d)$ *vérifie* $div(\mathbf{q}) = 0$.

Alors il existe $u \in L^\infty(0,T; H^1(\Omega))$ *tel que l'unique solution* u_δ *obtenue avec l'extension du schéma de Scharfetter-Gummel proposée satisfasse quand* $\delta \to 0$, *à l'extraction de sous-suites près,*

$$u_\delta \to u \text{ dans } L^2(\Omega \times (0,T)) \text{ fortement,}$$
$$\nabla^\delta u_\delta \rightharpoonup \nabla u \text{ dans } \left(L^2(\Omega \times (0,T))\right)^d \text{ faiblement,}$$

où $\nabla^\delta u_\delta$ *est une approximation du gradient de* u *obtenue par le schéma numérique. De plus, la limite* u *est solution faible de l'équation* (2.17).

Pour démontrer ce résultat, nous établissons d'abord l'existence, l'unicité et la stabilité L^∞ de la solution du schéma numérique dans la proposition 3.3.1. Pour démontrer ce résultat, nous faisons en particulier l'hypothèse $\mathrm{div}(\mathbf{q}) = 0$, qui n'est pas vérifiée si l'on s'intéresse par exemple au système de dérive-diffusion (2.11), puisque dans ce cas $\mathrm{div}(\mathbf{q}) = \Delta V \neq 0$. Cependant, si l'on suppose le dopage C nul sur Ω, nous pouvons prouver la proposition 3.3.1 pour le système (2.11). Le résultat clé pour prouver la convergence du schéma est l'estimation $L^2(0,T;H^1)$ donnée dans la proposition 3.3.2. C'est ici que l'hypothèse 2 intervient de manière cruciale car notre preuve nécessite d'avoir une borne uniforme inférieure strictement positive pour u_δ. Cette hypothèse implique en particulier que le problème ne dégénère pas. La compacité forte de u_δ est ensuite démontrée classiquement (proposition 3.4.1) en utilisant la continuité des translatées en temps et en espace de u_δ et l'estimation $L^2(0,T;H^1)$, par application du théorème de Riesz-Fréchet-Kolmogorov. La dernière étape consiste à passer à la limite dans le schéma (théorème 3.4.1). Comme dans le cas linéaire, la difficulté est de récupérer séparément à la limite les parties convective et diffusive, puisqu'elle sont discrétisées dans un même terme avec notre extension du flux de Scharfetter-Gummel.
Dans la dernière section du chapitre 3, nous appliquons le schéma au système de dérive-diffusion non linéaire (2.11) ainsi qu'à l'équation des milieux poreux (2.10). Nous constatons la décroissance exponentielle de l'entropie relative et de la dissipation d'entropie dans ces deux cas, sans le phénomène de saturation qui apparaît avec les schémas ne préservant pas les équilibres.

Construction d'un schéma d'ordre élevé

Dans les tests numériques du chapitre 3, nous ne considérons pas des situations trop fortement dégénérées. Si l'on considère des simulations avec des fonctions r dont la non-linéarité est plus importante ou des conditions de Dirichlet qui font dégénérer le problème

au bord, on constate une dégradation des résultats numériques obtenus avec notre extension du schéma de Scharfetter-Gummel. De plus, ce schéma ne permet pas de traiter le cas général où la convection est non linéaire (2.2).
Partant de ces constatations, nous proposons dans le chapitre 4 une nouvelle discrétisation en espace de l'équation (2.2). L'idée est toujours de proposer un flux numérique qui préserve les équilibres, mais qui reste de plus d'ordre élevé même dans le régime dégénéré. A cette fin, nous prenons en compte ensemble les termes de convection et de diffusion, en utilisant la formulation (2.4), qui permet de réécrire le flux sous la forme d'un flux d'advection :

$$f(u)\,\nabla V + \nabla r(u) = f(u)\,\nabla\left(V + h(u)\right),$$

puisque la fonction h est telle que $r'(s) = h'(s)\,f(s)$. Le terme $\nabla\left(V + h(u)\right)$ est alors considéré comme une « vitesse », dans laquelle contribuent à la fois la convection et la diffusion. On applique alors une méthode de discrétisation standard pour les équations de transport : le schéma de Lax-Friedrichs local, ou plus simplement le schéma décentré amont dans le cas linéaire $f(u) = u$. Ainsi, le flux numérique est défini de telle sorte que les états d'équilibres soient préservés et qu'un analogue discret de l'inégalité d'entropie–dissipation (2.5) puisse être automatiquement obtenu. Le schéma ainsi construit est alors seulement d'ordre un en espace. Il suffit ensuite d'appliquer une méthode de limiteurs de pente pour obtenir un schéma qui reste précis à l'ordre deux en espace, même dans le cas dégénéré.
Dans le chapitre 4, nous construisons en détail le schéma dans le cas unidimensionnel. La généralisation au cas multidimensionnel pour des maillages cartésiens de Ω est alors directe, puisqu'il faut simplement utiliser la construction unidimensionnelle dans chacune des directions cartésiennes. Par contre, la construction sur des maillages non structurés s'avère plus délicate. Plus précisément, il est immédiat d'obtenir le schéma d'ordre un pour de tels maillages en suivant la même démarche qu'en une dimension d'espace ; la difficulté est d'obtenir une méthode d'ordre deux. Pour cela, on peut par exemple se référer aux travaux de L. J. Durlofsky, B. Engquist, S. Osher [76] et E. Godlewski, P.-A. Raviart [98]. L'étude d'un schéma d'ordre élevé est assez complexe ; on prouve dans notre cas la positivité de la solution numérique obtenue, ainsi qu'une estimation d'entropie semi-discrète en espace lorsque la convection est linéaire (propositions 4.3.1 et 4.3.2). La discrétisation en temps considérée est un schéma d'Euler explicite, puisque nous nous focalisons ici sur la discrétisation spatiale du problème. Notons qu'en considérant un schéma d'Euler implicite, nous obtenons très facilement une estimation d'entropie–dissipation entièrement discrète.
Une grande partie du chapitre 4 est ensuite consacrée à la mise en œuvre de notre nouveau schéma numérique et à sa comparaison avec d'autres schémas volumes finis. Ainsi, nous vérifions tout d'abord numériquement que notre schéma est d'ordre deux en espace, même dans le cas dégénéré. Nous constatons par la suite son efficacité pour préserver l'asymptotique en temps long en l'appliquant aux modèles présentés précédemment : système de dérive-diffusion non linéaire pour les semi-conducteurs (2.11), équation des milieux poreux (2.10), équation de type Fokker-Planck non linéaire pour les fermions et les bosons (2.15).

3 Un schéma pour un modèle de chimiotactisme

Dans la partie III de ce manuscrit, nous analysons un schéma volumes finis pour le modèle de Keller-Segel en dimension deux avec diffusion croisée proposé par S. Hittmeir et A. Jüngel dans [111]. Cette analyse, présentée dans le chapitre 6, est fondée sur des estimations a priori déduites d'une inégalité d'entropie discrète dont l'obtention nécessite l'utilisation de versions discrètes d'inégalités fonctionnelles de type Gagliardo-Nirenberg-Sobolev et Poincaré-Sobolev. Dans le chapitre 5, nous commençons donc par établir ces inégalités, en nous plaçant pour cela dans un contexte assez général, incluant notamment le cas de conditions aux limites mixtes et une généralisation au cadre des schémas volumes finis en dualité discrète (ou *discrete duality finite volume*, abrégé par DDFV).

3.1 Présentation du modèle de chimiotactisme considéré

Dans le chapitre 6, nous nous intéressons donc à un schéma volumes finis pour un modèle de chimiotactisme en deux dimensions avec un terme de diffusion croisée additionnel dans l'équation sur la concentration en substance chimique. Commençons par rappeler ici quelques résultats concernant le modèle classique de Patlak-Keller-Segel, avant d'introduire le modèle modifié que nous considérons.

Le modèle de Patlak-Keller-Segel

Le chimiotactisme est le phénomène biologique par lequel des cellules, des bactéries ou d'autres organismes uni- ou pluricellulaires dirigent leurs mouvements en fonction de la concentration de certaines espèces chimiques présentes dans leur environnement. Ce phénomène joue un rôle important dans plusieurs domaines de la biologie, tels que l'embryogenèse, l'immunologie, la croissance tumorale ou encore la cicatrisation. Pour une description plus précise de ces modèles issus de la biologie, on pourra se reporter aux travaux de T. Hillen, K. Painter [110] et de B. Perthame [151].

Au niveau mascroscopique, les modèles de chimiotactisme prennent en compte deux entités : la densité de cellules $n(x,t)$ et la concentration en chimioattractant $S(x,t)$. Un modèle classique pour décrire l'évolution de ces deux variables est le système de Patlak-Keller-Segel, introduit par C. Patlak en 1953 [149] et par E. Keller et L. Segel en 1970 [127] :

$$\left\{ \begin{array}{rcl} \partial_t n & = & \operatorname{div}(\nabla n - n\,\nabla S), \\ \alpha\,\partial_t S & = & \Delta S + \mu n - S, \end{array} \right. \qquad x \in \Omega, \quad t \geq 0, \tag{3.1}$$

où $\Omega \subset \mathbb{R}^d$, $d \geq 1$, est un domaine borné ou l'espace $\mathbb{R}^d$ tout entier. Le paramètre $\mu > 0$ est le taux de sécrétion auquel la substance chimique est émise par les cellules. Le terme non linéaire $n\,\nabla S$ modélise le mouvement des cellules vers les zones où la concentration en substance chimique est la plus importante. Ce système est complété par des conditions aux limites de Neumann homogènes si Ω est un domaine borné et par des conditions initiales $n_0(x)$ et $S_0(x)$. La condition initiale sur S est nécessaire uniquement si $\alpha \neq 0$. Le paramètre $\alpha \in \{0,1\}$ mesure le rapport entre les échelles de temps de l'évolution en

concentration de chimioattractant et du mouvement des cellules. Quand $\alpha = 1$, le système (3.1) est de type parabolique-parabolique, et quand $\alpha = 0$ de type parabolique-elliptique. Le choix $\alpha = 0$ correspond à faire l'hypothèse que la concentration en chimioattractant évolue selon une échelle de temps beaucoup plus petite que la densité de cellules.

Ce modèle met en évidence un phénomène d'agrégation des cellules : plus les cellules sont regroupées, plus la concentration en substance chimique produite est importante. Ce phénomène est contrebalancé par la diffusion des cellules mais si la densité de ces dernières est suffisamment importante, les interactions chimiques dominent la diffusion et on observe alors une explosion de la densité de cellules. Pour le modèle en dimension 1, il n'y a pas de phénomène de seuil critique : l'existence de solutions globales en temps et bornées est prouvée (voir par exemple l'article de K. Osaki et A. Yagi [148]). En dimension 2, dans le cas parabolique-elliptique ($\alpha = 0$), T. Nagai [142] prouve que le seuil critique pour l'explosion est donné par $M = \int_\Omega n_0(x)\,dx = 4\pi$ si Ω est un domaine borné connexe de frontière $\mathcal{C}^2$. Dans le cas de données initiales ayant une symétrie radiale, T. Nagai, T. Senba et K. Yoshida [143] obtiennent un seuil de $M = 8\pi$, qui est aussi le seuil critique dans le cas où $\Omega = \mathbb{R}^2$ est l'espace tout entier, ainsi que l'ont démontré A. Blanchet, E. Carlen et J. A. Carrillo [22]. L'existence et l'unicité de solutions globales régulières dans le cas sous-critique est prouvée par W. Jäger et S. Luckhaus pour les domaines bornés [116] et par A. Blanchet, J. Dolbeault et B. Perthame pour l'espace entier $\mathbb{R}^2$ [24]. Dans le cas critique $M = 8\pi$ pour $\Omega = \mathbb{R}^2$, A. Blanchet, J. A. Carrillo et N. Masmoudi [23] démontrent l'existence d'une solution qui devient non bornée quand $t \to \infty$. De plus, M. Herrero et J. Velázquez [109] prouvent l'existence d'une donnée initiale à symétrie radiale pour laquelle, dans le cas sur-critique, la solution forme une δ–singularité en temps fini. Enfin, concernant le modèle parabolique-parabolique, les résultats obtenus par V. Calvez et L. Corrias [41] suggèrent dans le cas $\Omega = \mathbb{R}^2$ une masse critique de 8π, même si la preuve de l'explosion des solutions dans le cas sur-critique reste une question ouverte.
Des résultats ont également été obtenus pour le modèle en dimension $d \geq 3$. M. A. Herrero, E. Medina et J. L. L. Velázquez prouvent ainsi l'existence de solutions radiales qui explosent en temps fini pour le modèle parabolique-elliptique tridimensionnel [108], et M. Brenner, P. Constantin, L. P. Kadanoff, A. Schenkel et S. C. Venkataramani décrivent le comportement de solutions radiales du système parabolique-elliptique, qui dépend fortement de la dimension d de l'espace [29]. Concernant le système parabolique-parabolique, L. Corrias et B. Perthame [60] mettent en évidence le fait que le seuil critique pour l'explosion en dimension $d \geq 3$ est donné en terme de norme $L^{d/2+\varepsilon}$, $\varepsilon > 0$, pour la densité n et L^d pour la concentration S dans le cas où $\Omega = \mathbb{R}^d$, et M. Winkler [165] s'intéresse à cette question dans le cas d'un domaine borné. Pour une revue plus précise de résultats, nous renvoyons par exemple aux articles de D. Horstmann [112, 113].

Le modèle avec diffusion croisée

Un certain nombre de modifications du modèle de Keller-Segel (3.1) permettant d'éviter l'explosion en temps fini de la densité de cellules a été proposé ces dernières années, pour répondre à la fois à des problèmes de modélisation et à des difficultés numériques.

Une première idée consiste à modifier la sensibilité chimiotactique. Dans ce cas, l'existence globale de solutions peut être prouvée (voir par exemple l'article de M. Burger, M. Di Francesco et Y. Dolak-Struss [37] dans le cas parabolique-elliptique et l'article de M. Di Francesco et J. Rosado [71] pour le modèle parabolique-parabolique). V. Calvez et J. A. Carrillo [40] ainsi que R. Kowalczyk [128] considèrent quant à eux des termes de diffusion non linéaires, qui tiennent compte du fait que les cellules ont un certain volume. Une troisième méthode, adoptée par exemple par M. Winkler [164] revient à ajouter un terme de réaction dans l'équation sur la densité de cellules permettant de prendre en compte les naissances et les morts dans la population cellulaire.
Récemment, J. A. Carrillo, S. Hittmeir et A. Jüngel [44, 111] ont proposé une autre idée, qui consiste à introduire un terme additionnel de diffusion cellulaire dans l'équation sur la concentration en substance chimique. Ce modèle, duquel nous étudions une discrétisation dans le chapitre 6, s'écrit :

$$\left\{ \begin{array}{rcl} \partial_t n & = & \mathrm{div}(\nabla n - n\,\nabla S), \\ \alpha\,\partial_t S & = & \Delta S + \delta\,\Delta n + \mu n - S, \end{array} \right. \qquad x \in \Omega, \quad t \geq 0, \tag{3.2}$$

où $\delta > 0$ est la constante de diffusion additionnelle. S. Hittmeir et A. Jüngel [111] prouvent que ce terme de diffusion empêche l'explosion et conduit à l'existence de solutions faibles globales, même pour des constantes de diffusion arbitrairement petites. Une autre motivation pour introduire ce terme de diffusion croisée est d'essayer d'obtenir numériquement des estimations sur les temps d'explosion pour le modèle classique.
À première vue, l'ajout du terme de diffusion croisée $\delta\,\Delta n$ semble compliquer l'analyse mathématique. En effet, la matrice de diffusion du nouveau système (3.2) n'est plus symétrique ni définie positive, ce qui empêche d'appliquer le principe du maximum à l'équation sur la concentration en chimioattractant. Cependant, il est prouvé dans [111] que ces difficultés peuvent être contournées grâce au fait que le système modifié (3.2) possède une fonctionnelle d'entropie :

$$E(t) = \int_\Omega \left(n\,(\log(n) - 1) + \frac{\alpha}{2\,\delta}\,S^2 \right) dx, \tag{3.3}$$

qui satisfait l'équation de dissipation d'entropie suivante :

$$\frac{dE}{dt} + \int_\Omega \left(4\,|\nabla\sqrt{n}|^2 + \frac{1}{\delta}\,|\nabla S|^2 + \frac{1}{\delta}\,S^2 \right) dx = \frac{\mu}{\delta} \int_\Omega n\,S\,dx. \tag{3.4}$$

L'application d'inégalités fonctionnelles de type Gagliardo-Nirenberg-Sobolev et Poincaré-Sobolev au membre de droite de cette équation permet d'obtenir des estimations du gradient de $\sqrt{n}$ et de S, qui sont le point de départ pour analyser l'existence et le comportement en temps long des solutions. Plus précisément, en utilisant l'inégalité de Hölder et l'injection de Sobolev $H^1(\Omega) \hookrightarrow L^6(\Omega)$ en dimension 2 ou 3, on a :

$$\mu \int_\Omega n\,S\,dx \leq \mu\,C\,\|\,n\,\|_{L^{6/5}(\Omega)}\,\|\,S\,\|_{L^6(\Omega)} \leq \mu\,C\,\|\,\sqrt{n}\,\|^2_{L^{12/5}(\Omega)}\,\|\,S\,\|_{H^1(\Omega)}.$$

L'inégalité de Gagliardo-Nirenberg-Sobolev appliquée au terme en $\sqrt{n}$ avec $\theta = d/12$ (voir p. 125 dans [144]) et l'inégalité de Young impliquent finalement :

$$\begin{aligned}\mu \int_\Omega n\, S\, dx &\leq \mu\, C \,\|\sqrt{n}\,\|^{2\theta}_{H^1(\Omega)} \,\|\sqrt{n}\,\|^{2(1-\theta)}_{L^2(\Omega)} \,\|\, S\,\|_{H^1(\Omega)} \\ &\leq \mu\, C \,\|\sqrt{n}\,\|^{2\theta}_{H^1(\Omega)} \,\|\, n\,\|^{(1-\theta)}_{L^1(\Omega)} \,\|\, S\,\|_{H^1(\Omega)} \\ &\leq 2\,\delta \,\|\sqrt{n}\,\|^2_{L^2(\Omega)} + \frac{1}{2}\|\, S\,\|^2_{H^1(\Omega)} + C\left(\delta\,, \|\, n_0\,\|_{L^1(\Omega)}\right).\end{aligned}$$

Les deux premiers termes peuvent alors être pris en compte dans le membre de gauche de l'inégalité (3.4), tandis que le troisième terme est une constante dépendant uniquement de la dimension d, du paramètre δ et de la masse $\|\, n_0\,\|_{L^1(\Omega)}$ qui est une quantité conservée au cours du temps.
Dans le chapitre 6, notre but est d'analyser un schéma volumes finis pour le modèle de Keller-Segel modifié (3.2). Pour cela, nous adaptons au contexte discret les techniques utilisées dans le cadre continu, ce qui nécessite en particulier des versions discrètes des inégalités de Gagliardo-Nirenberg-Sobolev et de Poincaré-Sobolev employées ci-dessus.

3.2 Quelques inégalités fonctionnelles discrètes

Dans le chapitre 5, nous établissons donc quelques inégalités fonctionnelles discrètes qui sont souvent utiles pour analyser la convergence de schémas volumes finis.

État de l'art

Au niveau continu, les inégalités de Gagliardo-Nirenberg-Sobolev et de Poincaré-Sobolev constituent un outil fondamental pour l'analyse des équations aux dérivées partielles. Nous commençons par rappeler ces inégalités. Soit Ω un domaine ouvert borné de $\mathbb{R}^d$, avec $d \geq 2$.

– INÉGALITÉ DE GAGLIARDO-NIRENBERG-SOBOLEV [90, 144]. Pour tous $1 \leq p,\, q \leq \infty$, il existe une constante $C > 0$ telle que pour tout $u \in W^{1,p}(\Omega) \cap L^q(\Omega)$,

$$\|\, u\,\|_{L^m(\Omega)} \,\leq\, C\,\|\, u\,\|^{\theta}_{W^{1,p}(\Omega)}\,\|\, u\,\|^{1-\theta}_{L^q(\Omega)}, \tag{3.5}$$

avec

$$0 \leq \theta \leq 1 \quad \text{et} \quad \frac{1}{m} = \frac{1-\theta}{q} + \frac{\theta}{p} - \frac{\theta}{N}.$$

– INÉGALITÉ DE POINCARÉ-SOBOLEV [1, 31]. Pour tout $1 \leq p < \infty$, il existe une constante $C > 0$ telle que pour tout $u \in W^{1,p}(\Omega)$,

$$\|\, u\,\|_{L^q(\Omega)} \,\leq\, C\,\|\, u\,\|_{W^{1,p}(\Omega)}, \tag{3.6}$$

pour

$$1 \leq q \leq \frac{p\,d}{d-p} \quad \text{si} \quad 1 \leq p < d,$$

$$1 \leq q < \infty \qquad \text{si} \quad p \geq d.$$

Ces inégalités sont un outil standard pour l'étude de l'existence et de la régularité de solutions d'équations aux dérivées partielles elliptiques ou paraboliques. Le cadre L^2 est généralement utilisé pour les problèmes elliptiques linéaires, afin de montrer la coercivité de formes bilinéaires sur H_0^1 et ainsi de pouvoir appliquer le théorème de Lax-Milgram pour prouver l'existence de solutions faibles. Le cadre L^p est quant à lui fondamental pour l'étude d'équations elliptiques ou paraboliques non linéaires, afin d'obtenir des estimations d'énergie permettant de montrer l'existence de solutions faibles, comme nous venons par exemple de le voir dans la sous-section précédente 3.1.

L'analyse de la convergence et les estimations d'erreurs des méthodes numériques nécessitent très souvent l'emploi d'inégalités fonctionnelles discrètes. Plusieurs inégalités de type Poincaré-Sobolev ont été démontrées, aussi bien dans le cadre des schémas volumes finis que dans celui des discrétisations par éléments finis.
Concernant le contexte général des volumes finis, la première idée employée, pour prouver ces inégalités, a été d'utiliser des propriétés géométriques du maillage. Plus précisément, étant donnée une direction orientée $\mathcal{D}$, chaque centre de cellule du maillage est relié au centre d'une arête amont (par rapport à $\mathcal{D}$) du bord du domaine par une droite de direction $\mathcal{D}$. Cette droite coupe un certain nombre de cellules et leurs interfaces ; cet argument permet de relier la norme de la fonction continue par morceaux considérée à une norme d'une version discrète de son gradient. Pour des conditions au bord de Dirichlet, les premiers résultats ont été obtenus par Y. Coudière, J.-P. Vila et P. Villedieu [63], puis généralisés dans le livre de R. Eymard, T. Gallouët et R. Herbin [85] ainsi que dans les articles de Y. Coudière, J. Droniou, T. Gallouët et R. Herbin [62, 75]. Concernant le cas de conditions de Neumann, une inégalité de Poincaré-Wirtinger discrète a été établie par R. Eymard, T. Gallouët, R. Herbin et M.-H. Vignal [85, 93], toujours par la même méthode. Des inégalités de Sobolev plus générales pour des conditions de Neumann homogènes ont aussi été prouvées par C. Chainais-Hillairet et J. Droniou [50].
Plus récemment, une autre idée a été utilisée pour montrer ce type d'inégalités discrètes : l'injection continue de l'espace des fonctions à variations bornées $BV(\Omega)$ dans $L^{d/(d-1)}(\Omega)$, pour un domaine lipschitzien Ω. Cet argument a d'abord été employé par F. Filbet [89] pour prouver une inégalité de Poincaré-Sobolev discrète en dimension $d = 2$ dans le cas de conditions de Neumann homogènes. R. Eymard, T. Gallouët et R. Herbin [86] utilisent également cette méthode pour prouver des inégalités de Poincaré-Sobolev en toute dimension $d \geq 1$, dans le cas particulier de conditions de Dirichlet homogènes. Ce résultat est adapté au cas de conditions de Neumann et au cas de conditions aux limites mixtes par B. Andreianov, M. Bendahmane et R. Ruiz Baier [4]. Nous mentionnons par ailleurs l'article de F. Bouchut, R. Eymard et A. Prignet [25] où l'injection continue de $BV(\mathbb{R}^d)$ dans $L^{d/(d-1)}(\mathbb{R}^d)$ est utilisée pour obtenir une sorte d'inégalité de Gagliardo-Nirenberg-Sobolev dans l'espace entier $\mathbb{R}^d$, pour $d \geq 1$.
Finalement pour $p = 2$, A. Glitzky et J. A. Griepentrog [97] obtiennent des inégalités de Poincaré-Sobolev discrètes pour des approximations volumes finis sur des maillages de Voronoï, dans le cas de conditions aux limites quelconques. Leur preuve est une adaptation au cadre discret de celle proposée par Sobolev, qui est fondée sur une formule de représen-

tation intégrale, et qui utilise aussi fortement les propriétés de Voronoï du maillage.
Concernant à présent les discrétisations de type éléments finis, des variante de l'inégalité de Poincaré dans des espaces de Sobolev « brisés »ont été proposées par D. N. Arnold [12], S. C. Brenner [30] et M. Vohralik [163]. Un résultat généralisé a également été démontré par A. Lasis et E. Süli [131], en utilisant des résultats de régularité elliptique et des opérateurs d'interpolation pour des éléments finis non conformes. Finalement, D. A. Di Pietro et A. Ern [72] obtiennent un résultat dans le cadre non hilbertien, en s'inspirant de la technique utilisée par F. Filbet [89] et R. Eymard, T. Gallouët, R. Herbin [86], à savoir l'injection continue de $BV(\Omega)$ dans $L^{d/(d-1)}(\Omega)$.

Dans le chapitre 5, notre but est de donner une preuve simple de versions discrètes d'inégalités de type Gagliardo-Nirenberg-Sobolev (3.5) et Poincaré-Sobolev (3.6) pour des fonctions issues de discrétisations volumes finis, avec des conditions au bord arbitraires. Comme nous venons de le présenter, de nombreuses inégalités de Poincaré-Sobolev ont déjà été prouvées, mais nous tentons ici de donner des résultats unifiés, incluant en particulier le cas de conditions aux limites mixtes. Concernant les inégalités de type Gagliardo-Nirenberg-Sobolev, le seul résultat disponible à notre connaissance est celui de F. Bouchut, R. Eymard et A. Prignet [25], et il concerne l'espace tout entier $\mathbb{R}^d$ et non un domaine borné.
Comme dans [25, 72, 86, 89], notre point de départ est l'injection continue de $BV(\Omega)$ dans $L^{d/(d-1)}(\Omega)$. La principale difficulté apparaît lors de la prise en compte des conditions aux limites. Dans les travaux mentionnés précédemment [72, 86, 89], les conditions considérées sont les mêmes sur toute la frontière $\partial\Omega$, de type Dirichlet ou Neumann homogènes. Dans [25], le problème étant considéré dans $\mathbb{R}^d$ tout entier, il n'y a pas de conditions aux limites. Dans le cas où la fonction satisfait des conditions de Dirichlet homogènes seulement sur une partie $\Gamma^0 \subsetneq \partial\Omega$ de la frontière, nous ne pouvons pas appliquer la même stratégie que R. Eymard, T. Gallouët et R. Herbin [86], qui consiste à prolonger par zéro la fonction considérée à $\mathbb{R}^d$. Notre idée est alors d'épaissir la frontière de Ω pour prendre en compte les conditions aux limites mixtes dans ce cas.

Enfin, nous proposons une extension de ces résultats au cas d'approximations obtenues par des méthodes DDFV. Avant de présenter plus en détail nos résultats, dressons un bref historique concernant ces méthodes, qui ont été développées depuis une dizaine d'années pour approcher des problèmes elliptiques anisotropes sur des maillages généraux en 2D et 3D. Elles sont fondées sur des versions discrètes des opérateurs divergence et gradient, qui satisfont une formule de Green discrète. Ces approximations DDFV ont d'abord été proposées pour discrétiser des problèmes de diffusion anisotropes ou non linéaires, sur des maillages très généraux. Nous renvoyons ici aux travaux précurseurs de F. Hermeline [103, 104, 105] qui propose une nouvelle approche fondée sur l'utilisation de deux maillages, un maillage primal et un maillage dual, ainsi qu'à l'article de Y. Coudière, J.-P. Vila et P. Villedieu [63] qui présentent une méthode de reconstruction des gradients discrets. Ensuite, S. Delcourte, K. Domelevo et P. Omnès [69, 73] ont étudié une approche DDFV pour l'opérateur de Laplace. B. Andreianov, F. Boyer et F. Hubert [5] ont également

donné un cadre général pour les méthodes DDFV appliquées à des problèmes anisotropes et non linéaires. La plupart de ces travaux concernent des problèmes de diffusion linéaire 2D, anisotropes et hétérogènes. F. Boyer et F. Hubert [27] se sont alors intéressés au cas d'opérateurs de diffusion discontinus. Nous citons enfin les articles de F. Hermeline [106, 107] qui traitent de problèmes analogues en 3D, de S. Krell [129] concernant le problème de Stokes en dimensions 2 et 3, ainsi que de Y. Coudière et G. Manzini [61] qui s'intéressent à des équations stationnaires linéaires de convection-diffusion.

Présentation des résultats du chapitre 5

Étant donné un maillage admissible $\mathfrak{M}$ de Ω, nous notons $X(\mathfrak{M})$ l'ensemble des fonctions $u \in L^1(\Omega)$ constantes par mailles de $\mathfrak{M}$, et nous définissons des normes L^p et $W^{1,p}$ discrètes sur cet ensemble. Pour prouver les inégalités discrètes qui nous intéressent, nous utilisons des résultats concernant l'espace $BV(\Omega)$, muni de la norme

$$\| u \|_{BV(\Omega)} = \| u \|_{L^1(\Omega)} + TV_\Omega(u),$$

où $TV_\Omega(u)$ est la variation totale de u, semi-norme sur $BV(\Omega)$ définie par :

$$TV_\Omega(u) = \sup \left\{ \int_\Omega u(x) \operatorname{div}(\phi(x))\, dx, \quad \phi \in \mathcal{C}_c^1(\Omega), \quad |\phi(x)| \leq 1, \quad \forall x \in \Omega \right\}.$$

On pourra se référer aux livres de L. Ambrosio, N. Fusco, D. Pallara [3] et W. P. Ziemer [167] pour plus de détails sur l'espace $BV(\Omega)$.
Notre point de départ est synthétisé dans le théorème suivant (Theorem 5.2.1 et Theorem 5.2.8 du chapitre 5) :

Théorème 6. *Soit Ω un ouvert borné lipschitzien de $\mathbb{R}^d$, $d \geq 2$.*

1. *Il existe une constante $c(\Omega) > 0$ telle que*

$$\| u \|_{L^{d/(d-1)}(\Omega)} \leq c(\Omega) \| u \|_{BV(\Omega)} \quad \forall u \in BV(\Omega). \tag{3.7}$$

Si l'on suppose de plus que l'ouvert Ω est connexe, la semi-norme TV_Ω devient une norme sur l'espace des fonctions de $BV(\Omega)$ de moyenne nulle et sur l'espace des fonctions de $BV(\Omega)$ s'annulant sur une partie de la frontière $\partial\Omega$, et on a alors les résultats plus précis suivants.

2. *Il existe une constante $c(\Omega) > 0$ telle que*

$$\| u - \overline{u} \|_{L^{d/(d-1)}(\Omega)} \leq c(\Omega)\, TV_\Omega(u) \quad \forall u \in BV(\Omega), \tag{3.8}$$

où

$$\overline{u} = \frac{1}{m(\Omega)} \int_\Omega u(x)\, dx.$$

3. *Soit $\Gamma^0 \subset \partial\Omega$, $m(\Gamma^0) > 0$. Il existe une constante $c(\Omega) > 0$ dépendant uniquement de Ω telle que pour tout $u \in BV(\Omega)$ satisfaisant $u = 0$ sur Γ^0,*

$$\| u \|_{L^{d/(d-1)}(\Omega)} \leq c(\Omega)\, TV_\Omega(u). \tag{3.9}$$

En utilisant ce théorème et l'inclusion de $X(\mathfrak{M})$ dans $L^1(\Omega) \cap BV(\Omega)$, avec le contrôle de la norme $BV(\Omega)$ de $u \in X(\mathfrak{M})$ par la norme $W^{1,1}$ discrète, nous obtenons un lien entre la norme $L^{d/(d-1)}$ et la norme $W^{1,1}$, qui est le point de départ pour prouver les résultats du chapitre 5.

Dans l'ensemble du chapitre 5, nous supposons que Ω est un domaine ouvert borné polyédrique de $\mathbb{R}^d$, $d \geq 2$. Dans la section 5.3, nous ne prenons pas en compte les conditions au bord et nous prouvons des analogues discrets des inégalités de Gagliardo-Nirenberg-Sobolev (Theorem 5.3.1) et Poincaré-Sobolev (Theorem 5.3.2), en utilisant l'estimation (3.7). Ces inégalités sont utiles par exemple dans l'analyse de convergence de schémas volumes finis pour des problèmes avec des conditions aux limites de Neumann homogènes. Nous déduisons également des versions discrètes de l'inégalité de Nash :

$$\| u \|_{L^2(\Omega)}^{1+2/d} \leq C \| u \|_{L^1(\Omega)}^{2/d} \| u \|_{H^1(\Omega)} \quad \forall u \in L^1(\Omega) \cap H^1(\Omega),$$

et de l'inégalité de Poincaré-Wirtinger :

$$\| u - \overline{u} \|_{L^p(\Omega)} \leq C \| \nabla u \|_{L^p(\Omega)} \quad \forall u \in W^{1,p}(\Omega).$$

Pour montrer l'analogue discret de cette dernière inégalité, nous utilisons (3.8) au lieu de (3.7).

Dans la section 5.4, nous nous intéressons au cas plus délicat d'une approximation provenant d'un schéma volumes finis où des conditions au bord de Dirichlet homogènes sont prescrites sur une partie de mesure strictement positive de la frontière $\Gamma^0 \subset \partial\Omega$. Dans cette situation-là, nous prouvons des inégalités discrètes de type Gagliardo-Nirenberg-Sobolev (Theorem 5.4.1) et Poincaré-Sobolev (Theorem 5.4.2) où la norme $W^{1,p}$ apparaissant dans les résultats de la section 5.3 est remplacée par une semi-norme $W^{1,p}$ discrète qui prend en compte les sauts à la frontière et qui est en fait une norme sur l'espace des fonctions de $X(\mathfrak{M})$ s'annulant sur une partie de la frontière. Le point de départ est toujours l'injection continue de $BV(\Omega)$ dans $L^{d/(d-1)}(\Omega)$ donnée par (3.9) dans le cas de conditions de Dirichlet homogènes sur une partie de la frontière. Cependant, cette inégalité ne peut pas s'appliquer directement. En effet, si $u \in X(\mathfrak{M})$ est une fonction constante par mailles, elle appartient à $BV(\Omega)$ et sa trace sur la frontière est donc bien définie, mais elle ne s'annule pas forcément sur Γ^0. On doit donc adapter (3.9) pour en obtenir un analogue discret, ceci étant l'objet du lemme 5.4.1. Le principe est d'épaissir le domaine Ω en Ω_ε et de définir un prolongement u_ε de la fonction considérée u à Ω_ε tel que u_ε soit une fonction de $BV(\Omega_\varepsilon)$ et que la trace de u_ε s'annule sur une partie non vide de la frontière $\partial\Omega_\varepsilon$. Ceci permet d'appliquer (3.9) à u_ε et finalement d'obtenir l'analogue discret de (3.9) pour u en passant à la limite $\varepsilon \to 0$. Notons que pour prouver ce lemme, nous supposons que l'ouvert Ω est convexe.

Enfin dans la section 5.5 du chapitre 5, nous étendons les résultats des sections 5.3 et 5.4 à des approximations obtenues par des méthodes DDFV (Theorems 5.5.1, 5.5.2, 5.5.3 et 5.5.4).

3.3 Discrétisations du modèle de Keller-Segel

Avant de décrire les résultats obtenus dans le chapitre 6, présentons un rapide état de l'art des méthodes numériques proposées pour discrétiser le système de Keller-Segel classique (3.1).
Concernant le modèle parabolique-elliptique, des discrétisations par différences finies ont été proposées par N. Saito, T. Suzuki [157] et par R. Tyson, L. Stern, R. LeVeque [161]. A. Marrocco [139], N. Saito [155] ainsi que R. Strehl, A. Sokolov, D. Kuzmin et S. Turek [159] ont quant à eux utilisé des méthodes d'éléments finis. Citons également une méthode de maillages mobiles adaptatifs proposée par C. Budd, R. Carretero-González et R. Russell [35], à savoir un algorithme de la plus forte pente (*steepest descent approximation*) fondé sur le transport optimal introduit par A. Blanchet, V. Calvez et J. A. Carrillo [21], ainsi qu'une méthode particulaire stochastique étudiée par J. Haškovec et C. Schmeiser [101, 102].
Concernant les schémas numériques pour le modèle parabolique-parabolique, mentionnons la méthode volumes finis d'ordre deux d'A. Chertock et A. Kurganov [57], l'approche Galerkin discontinue proposée par Y. Epshteyn, A. Izmirlioglu et A. Kurganov [78, 79] et le schéma éléments finis introduit par N. Saito [156]. Citons également l'article de M. Burger, J. A. Carrillo et M.-T. Wolfram [36] pour une approximation par éléments finis mixtes d'un système de Keller-Segel avec diffusion non linéaire.
À notre connaissance, il existe peu de travaux dans lesquels l'analyse des schémas proposés est menée. F. Filbet [89] prouve l'existence et la convergence de solutions d'un schéma volumes finis pour le modèle parabolique-elliptique. Des estimations d'erreurs pour une approximation par éléments finis sont prouvées par N. Saito [155, 156]. Y. Epshteyn et A. Izmirlioglu prouvent quant à eux des estimations d'erreur pour une méthode de Galerkin discontinue [78]. Des preuves de convergence pour d'autres schémas numériques ont aussi été formulées par A. Blanchet, V. Calvez et J. A. Carrillo [21] ainsi que par J. Haškovec et C. Schmeiser [102].

3.4 Présentation des résultats du chapitre 6

Dans le chapitre 6, nous analysons un schéma volumes finis pour le modèle de Keller-Segel en dimension deux avec diffusion croisée (3.2), dans le cas parabolique-elliptique ($\alpha = 0$) :

$$\left\{ \begin{array}{rcl} \partial_t n & = & \mathrm{div}(\nabla n - n\,\nabla S), \\ 0 & = & \Delta S + \delta\,\Delta n + \mu n - S, \end{array} \right. \qquad x \in \Omega, \quad t \geq 0. \tag{3.10}$$

Nous imposons des conditions au bord de Neumann homogènes et une condition initiale pour n :

$$\nabla n \cdot \nu = \nabla S \cdot \nu = 0 \quad \text{sur } \partial\Omega, \quad t \geq 0, \tag{3.11}$$

$$n(x,0) = n_0(x), \quad x \in \Omega. \tag{3.12}$$

Nous considérons une discrétisation volumes finis en espace et une méthode d'Euler implicite en temps. Le schéma numérique est ici le même que celui proposé par F. Fil-

bet [89], mais avec un terme de diffusion croisée additionnel dans l'équation elliptique. Le schéma considéré étant implicite, nous commençons par établir l'existence d'une solution numérique. Nous prouvons également la préservation de la positivité et la conservation de la masse par le schéma ; c'est l'objet du Theorem 6.2.1. La preuve de ce théorème est réalisée de manière classique, en appliquant le théorème de Brouwer. L'opérateur de point fixe est défini à partir d'une linéarisation du schéma.
Nous étudions ensuite la convergence du schéma numérique vers la solution faible du système (3.10)–(3.12). Cette analyse est fondée sur des estimations a priori obtenues grâce à une inégalité d'entropie (Proposition 6.4.1) qui est l'analogue discret de l'équation (3.4) :

Proposition 7. *Il existe une constante $C > 0$ dépendant uniquement de Ω, μ, δ et n_0 telle que pour tout $k \geq 0$:*

$$E^{k+1} - E^k + \Delta t\, \mathcal{I}^{k+1} \leq C\, \Delta t, \tag{3.13}$$

où E^k est une approximation de l'entropie (3.3) au temps $t^k = k\,\Delta t$, et $\mathcal{I}^k$ est une approximation de la dissipation

$$\mathcal{I}(t) = \int_\Omega \left(4\left|\nabla \sqrt{n}\right|^2 + \frac{1}{\delta}\left|\nabla S\right|^2 + \frac{1}{\delta} S^2 \right) dx.$$

Cette inégalité d'entropie (3.13) est obtenue en utilisant des versions discrètes d'inégalités de Sobolev prouvées dans le chapitre 5.
En notant (n_η, S_η) la solution obtenue par le schéma pour une discrétisation espace-temps de taille η, nous obtenons, grâce à l'estimation d'entropie (3.13), les bornes suivantes, uniformes en η :

$$\begin{gathered}(n_\eta),\ (n_\eta \log(n_\eta)) \text{ sont bornées dans } L^\infty(0,T;L^1(\Omega)),\\ (dn_\eta) \text{ est bornée dans } L^2(\Omega\times(0,T)),\\ (n_\eta),\ (S_\eta) \text{ sont bornées dans } L^2(0,T;H^1(\Omega)),\end{gathered}$$

où dn_η est une approximation du gradient de n obtenue avec le schéma. Nous obtenons alors la compacité d'une famille de solutions approchées et nous pouvons passer à la limite dans le schéma (Theorem 6.2.2) :

Théorème 8. *Il existe n, $S \in L^2(0,T;H^1(\Omega))$ tels que la solution (n_η, S_η) obtenue avec le schéma numérique satisfasse, quand $\eta \to 0$, à l'extraction de sous-suites près,*

$$\begin{gathered} n_\eta \to n \quad \text{dans } L^2(\Omega\times(0,T)) \text{ fortement},\\ dn_\eta \rightharpoonup \nabla n, \quad S_\eta \rightharpoonup S, \quad dS_\eta \rightharpoonup \nabla S \quad \text{dans } L^2(\Omega\times(0,T)) \text{ faiblement}.\end{gathered}$$

De plus, la limite (n, S) est solution faible du système (3.10)–(3.12).

Nous nous intéressons également au comportement en temps long des solutions numériques. Au niveau continu, S. Hittmeir et A. Jüngel [111] prouvent que si le taux de sécrétion $\mu > 0$ est suffisamment petit ou si le paramètre de diffusion $\delta > 0$ est suffisamment

grand, alors la solution (n, S) de (3.10)–(3.12) converge vers l'état stationnaire homogène (n^*, S^*), où

$$n^* = \frac{\| n_0 \|_{L^1(\Omega)}}{\mathrm{m}(\Omega)} \quad \text{et} \quad S^* = \mu\, n^*.$$

En introduisant une version discrète de l'entropie relative définie par rapport à l'état stationnaire homogène

$$E[n(t)|n^*] = \int_\Omega n \log\left(\frac{n}{n^*}\right) dx, \tag{3.14}$$

et en montrant une estimation d'entropie–dissipation discrète, nous obtenons la convergence des solutions numériques vers l'état stationnaire homogène quand $t \to \infty$ si μ est suffisamment petit ou δ est suffisamment grand (Theorem 6.2.3).
Enfin, nous concluons le chapitre 6 en présentant quelques simulations numériques qui semblent indiquer l'existence de solutions stationnaires non homogènes pour des valeurs intermédiaires du paramètre de diffusion δ.

Première partie

Étude d'un schéma préservant des asymptotiques : le schéma de Scharfetter-Gummel tout implicite

Dans cette première partie, nous nous intéressons à la discrétisation du système de dérive-diffusion isotherme pour les semi-conducteurs :

$$\text{(0.3.15)} \qquad \begin{cases} \partial_t N - \operatorname{div}(\nabla N - N\nabla\Psi) = 0, \\ \partial_t P - \operatorname{div}(\nabla P + P\nabla\Psi) = 0, \\ -\lambda^2 \Delta\Psi = P - N - C, \end{cases}$$

où $N(x,t)$ désigne la densité d'électrons, $P(x,t)$ la densité de trous et $\Psi(x,t)$ le potentiel électrostatique qui sont les inconnues du problème. Le profil de dopage $C(x)$ est donné. Le paramètre λ, appelé longueur de Debye, provient de l'adimensionnement du modèle physique.
Nous considérons une discrétisation de ce problème par une méthode d'Euler implicite en temps et un schéma de type volumes finis en espace, où les flux numériques choisis sont ceux de Scharfetter-Gummel [114, 158]. Notre but dans cette partie est d'étudier la préservation de deux asymptotiques par ce schéma : d'une part l'asymptotique en temps long ($t \to +\infty$), et d'autre part la limite quasi-neutre ($\lambda \to 0$). Au niveau continu, la convergence en temps long à un taux exponentiel vers l'équilibre thermique a été étudiée dans l'article de H. Gajewski et K. Gärtner [92]. Concernant la limite quasi-neutre, elle a été analysée par A. Jüngel et Y.-J. Peng dans le cas de conditions aux limites mixtes [121] et par I. Gasser, C. D. Levermore, P. Markowich et C. Schmeiser dans le cas de conditions de Neumann homogènes [94, 95]. Pour l'étude de ces deux asymptotiques, le point clé des preuves est une estimation d'énergie avec contrôle de la dissipation d'énergie. L'idée est d'adapter ces techniques de preuves au niveau discret.

Dans le chapitre 1, nous étudions d'abord le comportement en temps long du schéma de Scharfetter-Gummel. Plus précisément, nous prouvons la convergence en temps long de la solution approchée du système (0.3.15) obtenue avec le schéma de Scharfetter-Gummel implicite vers une approximation de l'équilibre thermique calculée avec le schéma proposé dans [51]. Le point clé de la preuve est une estimation d'énergie discrète qui permet de contrôler la dissipation d'énergie. De plus, les simulations numériques effectuées montrent que la vitesse de convergence vers l'équilibre thermique est bien exponentielle.

Dans le chapitre 2, nous nous intéressons à la préservation au niveau discret de la limite quasi-neutre dans le système de dérive-diffusion (0.3.15), toujours par le schéma de Scharfetter-Gummel implicite. En partant à nouveau d'une estimation d'énergie discrète proche de celle prouvée dans le chapitre 1, nous obtenons des estimations a priori indépendantes de la longueur de Debye, qui permettent de prouver la convergence du schéma vers une solution faible de (0.3.15) pour tout $\lambda > 0$.

Chapitre 1

Comportement en temps long du schéma de Scharfetter-Gummel pour le modèle de dérive-diffusion *

Dans ce chapitre, nous étudions le comportement en temps long du schéma de Scharfetter-Gummel implicite pour le modèle de dérive-diffusion linéaire pour les semi-conducteurs. Nous prouvons la convergence des solutions numériques vers une approximation de l'équilibre thermique en établissant une inégalité d'entropie–dissipation d'entropie discrète. Nous présentons également des simulations numériques qui soulignent la préservation du comportement en temps long des solutions approchées obtenues.

∗. Ce chapitre reprend les résultats présentés dans *Asymptotic Behavior of the Scharfetter-Gummel Scheme for the Drift-Diffusion Model* [55], publié dans les *proceedings* du congrès FVCA VI, 6th International Symposium on Finite Volume for Complex Applications. Nous mentionnons l'ajout de la preuve d'existence de solutions au schéma tout implicite (Theorem 1.2.1).

Contents

1.1 Introduction

1.1.1 Presentation of the problem

In the modeling of semiconductor devices, there is a hierarchy of models ranging from the kinetic transport equations to the drift-diffusion equations. In this hierarchy, the drift-diffusion system is widely used as it simplifies computations while giving an accurate description of the device physics.
Let $\Omega \subset \mathbb{R}^d$ $(d \geq 1)$ be an open and bounded domain describing the geometry of the semiconductor device. The isothermal drift-diffusion system consists of two continuity equations for the electron density $N(x,t)$ and the hole density $P(x,t)$, and a Poisson equation for the electrostatic potential $\Psi(x,t)$:

$$\left\{\begin{array}{lll} \partial_t N - \operatorname{div}(\nabla N - N\nabla\Psi) = 0 & \text{on} & \Omega\times(0,T), \\ \partial_t P - \operatorname{div}(\nabla P + P\nabla\Psi) = 0 & \text{on} & \Omega\times(0,T), \\ \lambda^2\Delta\Psi = N - P - C & \text{on} & \Omega\times(0,T), \end{array}\right. \tag{1.1.1}$$

where $C(x)$ is the doping profile, which is assumed to be a given datum, and λ is the rescaled Debye length arising from the scaling of the physical model. We supplement these equations with initial conditions N_0 and P_0:

$$N(x,0) = N_0(x), \quad P(x,0) = P_0(x), \quad x \in \Omega, \tag{1.1.2}$$

and physically motivated boundary conditions: Dirichlet boundary conditions on ohmic contacts Γ^D:

$$N(\gamma,t) = N^D(\gamma,t),\ P(\gamma,t) = P^D(\gamma,t),\ \Psi(\gamma,t) = \Psi^D(\gamma,t),\ (\gamma,t) \in \Gamma^D\times[0,T], \tag{1.1.3}$$

and homogeneous Neumann boundary conditions on insulating boundary segments Γ^N:

$$\nabla N\cdot\nu = \nabla P\cdot\nu = \nabla\Psi\cdot\nu = 0 \quad \text{on } \Gamma^N\times[0,T]. \tag{1.1.4}$$

There is an extensive literature on numerical schemes for the drift-diffusion equations: finite difference methods, finite elements methods, mixed exponential fitting finite elements methods, finite volume methods (see [32]). The Scharfetter–Gummel scheme is widely used to approximate the drift-diffusion equations in the linear case. It has been proposed and studied in [114] and [158]. It preserves steady-state, and is second order accurate in space (see [132]).
The purpose of this chapter is to study the large time behavior of the numerical solution given by the implicit Scharfetter-Gummel scheme for the transient linear drift-diffusion model (1.1.1). Indeed, it has been proved by H. Gajewski and K. Gärtner in [92] that the solution to the transient system (1.1.1) converges to the thermal equilibrium state as $t \to \infty$ if the boundary conditions are in thermal equilibrium. A. Jüngel extends this result to a degenerate model with nonlinear diffusivities in [119].
More precisely, the thermal equilibrium is a particular steady-state for which electron and hole currents, namely $\nabla N - N\nabla\Psi$ and $\nabla P + P\nabla\Psi$, vanish. If the Dirichlet boundary conditions satisfy N^D, $P^D > 0$ and

$$\log(N^D) - \Psi^D = \alpha_N \text{ and } \log(P^D) + \Psi^D = \alpha_P \text{ on } \Gamma^D, \tag{1.1.5}$$

the thermal equilibrium is defined by

$$\begin{cases} \lambda^2 \Delta \Psi^{eq} = \exp\left(\alpha_N + \Psi^{eq}\right) - \exp\left(\alpha_P - \Psi^{eq}\right) - C & \text{on } \Omega, \\ N^{eq} = \exp\left(\alpha_N + \Psi^{eq}\right), \ \ P^{eq} = \exp\left(\alpha_P - \Psi^{eq}\right) & \text{on } \Omega, \end{cases} \tag{1.1.6}$$

with the boundary conditions (1.1.3)–(1.1.4).
Long-time behavior of solutions to discretized drift-diffusion systems have been studied in [92], [51] and [96]. The proof of convergence to the thermal equilibrium is based on the following energy estimate:

$$0 \leq \mathcal{E}(t) + \int_0^t \mathcal{I}(\tau)\, d\tau \leq \mathcal{E}(0), \tag{1.1.7}$$

where $\mathcal{E}$ is the relative energy defined by

$$\begin{aligned} \mathcal{E}(t) = \int_\Omega \Big(& H\left(N(t)\right) - H\left(N^{eq}\right) - \log\left(N^{eq}\right)\left(N(t) - N^{eq}\right) \\ & + H\left(P(t)\right) - H\left(P^{eq}\right) - \log\left(P^{eq}\right)\left(P(t) - P^{eq}\right) \\ & + \frac{\lambda^2}{2}\left|\nabla\left(\Psi(t) - \Psi^{eq}\right)\right|^2 \Big)\, dx, \end{aligned} \tag{1.1.8}$$

with $H(s) = \int_1^s \log(\tau)\, d\tau$, and the energy dissipation $\mathcal{I}$ is given by

$$\mathcal{I}(t) = -\int_\Omega \left(N(t)\left|\nabla\left(N(t) - \Psi(t)\right)\right|^2 + P(t)\left|\nabla\left(P(t) + \Psi(t)\right)\right|^2 \right) dx. \tag{1.1.9}$$

Our aim is to prove that the solution of the Scharfetter–Gummel scheme converges to an approximation of the thermal equilibrium as $t \to +\infty$, and to this end we prove a discrete version of the energy estimate (1.1.7).
In the sequel, we will suppose that the following hypotheses are fulfilled:

(H1) N^D, P^D and Ψ^D are traces on $\Gamma^D \times (0,T)$ of functions, also denoted N^D, P^D and Ψ^D, such that $N^D,\ P^D \in H^1(\Omega\times(0,T)) \cap L^\infty(\Omega\times(0,T))$, $\Psi^D \in L^\infty\left(0,T;H^1(\Omega)\right)$ and $N^D,\ P^D \geq 0$ *a.e.*,

(H2) $N_0,\ P_0 \in L^\infty(\Omega)$ and $N_0,\ P_0 \geq 0$ *a.e.*,

(H3) there exist $0 < m \leq M$ such that: $m \leq N^D, N_0, P^D, P_0 \leq M$,

(H4) N^D, P^D and Ψ^D satisfy the compatibility condition (1.1.5).

1.1.2 Numerical schemes and main result

In this subsection, we first present the finite volume schemes for the time evolution drift-diffusion system (1.1.1) and for the thermal equilibrium (1.1.6).
An admissible mesh of Ω is given by a family $\mathcal{T}$ of control volumes (open and convex polygons in 2-D, polyhedra in 3-D), a family $\mathcal{E}$ of edges in 2-D (faces in 3-D) and a family of points $(x_K)_{K\in\mathcal{T}}$ which satisfy Definition 5.1 in [85]. It implies that the straight line between two neighboring centers of cells (x_K, x_L) is orthogonal to the edge $\sigma = K|L$.
In the set of edges $\mathcal{E}$, we distinguish the interior edges $\sigma \in \mathcal{E}_{int}$ and the boundary edges $\sigma \in \mathcal{E}_{ext}$. We split $\mathcal{E}_{ext}$ into $\mathcal{E}_{ext} = \mathcal{E}_{ext}^D \cup \mathcal{E}_{ext}^N$ where $\mathcal{E}_{ext}^D$ is the set of Dirichlet boundary edges and $\mathcal{E}_{ext}^N$ is the set of Neumann boundary edges. For a control volume $K \in \mathcal{T}$, we denote by $\mathcal{E}_K$ the set of its edges, which is also split into $\mathcal{E}_K = \mathcal{E}_{K,int} \cup \mathcal{E}_{K,ext}^D \cup \mathcal{E}_{K,ext}^N$.
We denote by d the distance in $\mathbb{R}^d$ and m the measure in $\mathbb{R}^d$ or $\mathbb{R}^{d-1}$. We also need some regularity assumption on the mesh:

$$\exists\, \xi > 0 \text{ s. t. } \mathrm{d}(x_K,\sigma) \geq \xi\, \mathrm{d}(x_K,x_L) \text{ for } K \in \mathcal{T},\ \text{for } \sigma = K|L \in \mathcal{E}_{int,K}. \tag{1.1.10}$$

For all $\sigma \in \mathcal{E}$, we define the transmissibility coefficient $\tau_\sigma = \mathrm{m}(\sigma)/d_\sigma$, where $d_\sigma = \mathrm{d}(x_K,x_L)$ for $\sigma = K|L \in \mathcal{E}_{int}$ and $d_\sigma = \mathrm{d}(x_K,\sigma)$ for $\sigma \in \mathcal{E}_{ext}$.
Let $(\mathcal{T},\mathcal{E},(x_K)_{K\in\mathcal{T}})$ be an admissible discretization of Ω and let us define the time step Δt, $N_T = E(T/\Delta t)$ and the increasing sequence $(t^n)_{0\leq n\leq N_T}$, where $t^n = n\Delta t$, in order to get a space-time discretization $\mathcal{D}$ of $\Omega\times(0,T)$.

First of all, the initial conditions and the doping profile are approximated by taking the mean values of N_0, P_0 and C on each cell K: for $U = \{N^0, P^0, C\}$,

$$U_K = \frac{1}{\mathrm{m(K)}} \int_K U(x)\, dx, \quad K \in \mathcal{T}.$$

The numerical boundary conditions $\left(N_\sigma^{n+1}, P_\sigma^{n+1}, \Psi_\sigma^{n+1}\right)_{n\geq 0,\sigma\in\mathcal{E}_{ext}^D}$ are also given by the mean values of (N^D, P^D, Ψ^D) on $\sigma\times[t^n, t^{n+1}[$: for $U = \{N, P, \Psi\}$

$$U_\sigma^{n+1} = \frac{1}{\Delta t\, \mathrm{m}(\sigma)} \int_{t^n}^{t^{n+1}} \int_\sigma U^D(\gamma,t)\, d\gamma\, dt, \quad \sigma \in \mathcal{E}_{ext}^D, \quad n \geq 0. \tag{1.1.11}$$

For any discrete quantity $\left((U_K)_{K\in\mathcal{T}}, (U_\sigma)_{\sigma\in\mathcal{E}_{ext}^D}\right)$, we also introduce the following notation: for all $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_{K,int} \cup \mathcal{E}_{K,ext}^D$,

$$U_{K,\sigma} = \begin{cases} U_L & \text{if} \quad \sigma = K|L \in \mathcal{E}_{K,int}, \\ U_\sigma & \text{if} \quad \sigma \in \mathcal{E}_{K,ext}^D. \end{cases}$$

Then for a given function f, for all $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$, we define

$$Df(U)_{K,\sigma} = \begin{cases} f(U_{K,\sigma}) - f(U_K) & \text{if} \quad \sigma \in \mathcal{E}_{K,int} \cup \mathcal{E}_{K,ext}^D, \\ 0 & \text{if} \quad \sigma \in \mathcal{E}_{K,ext}^N, \end{cases}$$

and

$$D_\sigma f(U) = |Df(U)_{K,\sigma}|\,.$$

The scheme for the thermal equilibrium

We compute an approximation $(N_K^{eq}, P_K^{eq}, \Psi_K^{eq})_{K\in\mathcal{T}}$ of the thermal equilibrium $(N^{eq}, P^{eq}, \Psi^{eq})$ defined by (1.1.6) with the finite volume scheme proposed by C. Chainais-Hillairet and F. Filbet in [51]:

(1.1.12)
$$\begin{cases} \lambda^2 \displaystyle\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma D\Psi_{K,\sigma}^{eq} = \mathrm{m}(K)\left(\exp(\alpha_N + \Psi_K^{eq}) - \exp(\alpha_P - \Psi_K^{eq}) - C_K\right) & \forall K \in \mathcal{T}, \\ N_K^{eq} = \exp(\alpha_N + \Psi_K^{eq}), \quad P_K^{eq} = \exp(\alpha_P - \Psi_K^{eq}) & \forall K \in \mathcal{T}. \end{cases}$$

Assuming that the boundary conditions satisfy hypotheses (H1)–(H4), the scheme (1.1.12) admits a unique solution (see [51, Proposition 3.1]).

The scheme for the transient model

The Scharfetter-Gummel scheme for the system (1.1.1) is defined by:

(1.1.13)
$$\begin{cases} \mathrm{m}(K)\dfrac{N_K^{n+1} - N_K^n}{\Delta t} + \displaystyle\sum_{\sigma\in\mathcal{E}_K} \mathcal{F}_{K,\sigma}^{n+1} = 0, & \forall K \in \mathcal{T}, \forall n \geq 0, \\ \mathrm{m}(K)\dfrac{P_K^{n+1} - P_K^n}{\Delta t} + \displaystyle\sum_{\sigma\in\mathcal{E}_K} \mathcal{G}_{K,\sigma}^{n+1} = 0, & \forall K \in \mathcal{T}, \forall n \geq 0, \\ \lambda^2 \displaystyle\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma D\Psi_{K,\sigma}^{n} = \mathrm{m}(K)\left(N_K^n - P_K^n - C_K\right), & \forall K \in \mathcal{T}, \forall n \geq 0, \end{cases}$$

with for all $\sigma \in \mathcal{E}_K$

(1.1.14)
$$\mathcal{F}_{K,\sigma}^{n+1} = \begin{cases} \tau_\sigma \left(B\left(-D\Psi_{K,\sigma}^{n+1}\right) N_K^{n+1} - B\left(D\Psi_{K,\sigma}^{n+1}\right) N_{K,\sigma}^{n+1}\right) & \text{if} \quad \sigma \in \mathcal{E}_{K,int} \cup \mathcal{E}_{K,ext}^D, \\ 0 & \text{if} \quad \sigma \in \mathcal{E}_{K,ext}^N, \end{cases}$$

(1.1.15)
$$\mathcal{G}_{K,\sigma}^{n+1} = \begin{cases} \tau_\sigma \left(B\left(D\Psi_{K,\sigma}^{n+1}\right) P_K^{n+1} - B\left(-D\Psi_{K,\sigma}^{n+1}\right) P_{K,\sigma}^{n+1}\right) & \text{if} \quad \sigma \in \mathcal{E}_{K,int} \cup \mathcal{E}_{K,ext}^D, \\ 0 & \text{if} \quad \sigma \in \mathcal{E}_{K,ext}^N, \end{cases}$$

where B is the Bernoulli function defined by:

$$B(x) = \frac{x}{e^x - 1} \text{ for } x \neq 0, \quad B(0) = 1. \tag{1.1.16}$$

These numerical fluxes have been introduced by A. M. Il'in in [114] and D. L. Scharfetter and H. K. Gummel in [158] for the numerical approximation of convection-diffusion terms with linear diffusion. It has been established by R. D. Lazarov, I. D. Mishev and P. S. Vassilevsky in [132] that they are second-order accurate in space. Moreover, they preserve steady-states.
We consider a fully implicit discretization in time to avoid the restrictive stability condition $\Delta t \leq \lambda^2/M$. In Section 1.2, we will establish the existence of a solution to this scheme, and L^∞ estimates on the approximate densities.

We may now state our main result, which is proved in Section 1.4:

Theorem 1.1.1. *Let us assume (H1)–(H4) and $C = 0$. Then solution $(N_\delta, P_\delta, \Psi_\delta)$ given by the scheme* (1.1.13)–(1.1.14)–(1.1.15) *satisfies for each $K \in \mathcal{T}$*

$$(N_K^n, P_K^n, \Psi_K^n) \longrightarrow (N_K^{eq}, P_K^{eq}, \Psi_K^{eq}) \text{ as } n \to +\infty,$$

where $(N_K^{eq}, P_K^{eq}, \Psi_K^{eq})_{K\in\mathcal{T}}$ is an approximation to the solution of the steady-state equation (1.1.6) given by (1.1.12).

The proof is based, as in the continuous case (see [92] and [119]), on an energy estimate and a control of its dissipation, given in Proposition 1.4.1 which is valid even if $C \neq 0$. Nevertheless to prove rigorously the convergence to equilibrium, we need the uniform lower bound (1.2.1) on N and P which holds under the restrictive assumption $C = 0$.

The plan of the chapter is as follows. In the next section, we prove the existence and the L^∞ stability of the approximate solutions of the scheme. In Section 1.3, we establish some technical properties of the Scharfetter-Gummel fluxes, which will be crucial in the proof of energy estimates in this chapter and in Chapter 2. Section 1.4 is devoted to the proof of Theorem 1.1.1 based on a discrete version of the energy estimate (1.1.7). Finally we perform some numerical experiments in the last section and observe the convergence to the equilibrium even when the doping profile C does not vanish.

1.2 Existence of a solution to the numerical scheme

In this section, we prove the existence of a solution to the numerical scheme. Indeed, as the scheme is fully implicit in time, we must prove that the nonlinear system of equations (1.1.13)–(1.1.14)–(1.1.15) admits a solution at each time step. Therefore, we will use a fixed-point theorem. Then, we also get L^∞ estimates on the approximate densities, which will be crucial to prove the convergence to the approximate thermal equilibrium.

Theorem 1.2.1. *Let us assume (H1)–(H2)–(H3) and that the doping profile vanishes on the whole domain: $C = 0$.*
Then there exists a solution $\{(N_K^n, P_K^n, \Psi_K^n),\ K \in \mathcal{T},\ 0 \le n \le N_T\}$ to the scheme (1.1.13)–(1.1.14)–(1.1.15)*, and the densities satisfy the following L^∞ estimate:*

$$0 < m \le N_K^n,\ P_K^n \le M, \quad \forall K \in \mathcal{T}, \quad \forall\, 0 \le n \le N_T. \tag{1.2.1}$$

Proof. The proof is based on the Brouwer fixed-point theorem and uses some ideas developed by A. Prohl and M. Schmuck in [153] to get the unconditionnal L^∞ stability of the scheme.
We prove the result by induction on $n \ge 0$. First, the result is clear for $n = 0$ by hypothesis (H3). We suppose now that for some $n \ge 0$, (N^n, P^n, Ψ^n) exists and satisfies (1.2.1). Then we will prove the existence of a solution $(N^{n+1}, P^{n+1}, \Psi^{n+1})$ to (1.1.13)–(1.1.14)–(1.1.15) such that

$$0 < m \le N_K^{n+1},\ P_K^{n+1} \le M \quad \forall K \in \mathcal{T}.$$

Let $\alpha > 0$ be a real number, which will be chosen later. We construct an application

$$\begin{array}{rccc} T_\alpha^n : & \mathbb{R}^\theta \times \mathbb{R}^\theta & \to & \mathbb{R}^\theta \times \mathbb{R}^\theta \\ & (N, P) & \mapsto & (\tilde{N}, \tilde{P}) \end{array}$$

based on a linearization of the scheme (1.1.13)–(1.1.14)–(1.1.15), where θ is the number of control volumes $K \in \mathcal{T}$.
For $(N, P) \in \mathbb{R}^\theta \times \mathbb{R}^\theta$, we solve the following linearized problem.

– We construct $\Psi \in \mathbb{R}^\theta$ using the following linear scheme: for all $K \in \mathcal{T}$,

$$\lambda^2 \sum_{\sigma \in \mathcal{E}_K} \tau_\sigma D\Psi_{K,\sigma} = \mathrm{m}(K)(N_K - P_K), \tag{1.2.2}$$

and $\Psi_\sigma = \Psi_\sigma^{n+1}$ defined by (1.1.11) for $\sigma \in \mathcal{E}_{ext}^D$. The existence and uniqueness of Ψ satisfying this system are obvious.

– Then we construct $(\tilde{N}, \tilde{P})$ using the following linear scheme: for all $K \in \mathcal{T}$,

$$\begin{aligned} &\frac{\mathrm{m}(K)}{\Delta t}\left(1 + \frac{\alpha}{\lambda^2}\right)\tilde{N}_K + \sum_{\sigma \in \mathcal{E}_K} \tau_\sigma \left(B\left(-D\Psi_{K,\sigma}\right)\tilde{N}_K - B\left(D\Psi_{K,\sigma}\right)\tilde{N}_{K,\sigma}\right) \\ &\qquad = \frac{\mathrm{m}(K)}{\Delta t} N_K^n + \frac{\mathrm{m}(K)}{\Delta t}\,\frac{\alpha}{\lambda^2}\, N_K, \end{aligned} \tag{1.2.3}$$

$$\begin{aligned} &\frac{\mathrm{m}(K)}{\Delta t}\left(1 + \frac{\alpha}{\lambda^2}\right)\tilde{P}_K + \sum_{\sigma \in \mathcal{E}_K} \tau_\sigma \left(B\left(D\Psi_{K,\sigma}\right)\tilde{P}_K - B\left(-D\Psi_{K,\sigma}\right)\tilde{P}_{K,\sigma}\right) \\ &\qquad = \frac{\mathrm{m}(K)}{\Delta t} P_K^n + \frac{\mathrm{m}(K)}{\Delta t}\,\frac{\alpha}{\lambda^2}\, P_K, \end{aligned} \tag{1.2.4}$$

and $\tilde{N}_\sigma = N_\sigma^{n+1}$, $\tilde{P}_\sigma = P_\sigma^{n+1}$ for $\sigma \in \mathcal{E}_{ext}^D$.

The scheme (1.2.3) can be written as $A_N\, \tilde{N} = S_N$, where:

– A_N is the sparse matrix defined by

$$(A_N)_{K,K} = \frac{\mathrm{m}(K)}{\Delta t}\left(1+\frac{\alpha}{\lambda^2}\right) + \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\, B\left(-D\Psi_{K,\sigma}\right) \quad \forall K\in\mathcal{T},$$
$$(A_N)_{K,L} = -\tau_\sigma\, B\left(D\Psi_{K,\sigma}\right) \quad \forall L\in\mathcal{T} \text{ such that } \sigma = K|L \in \mathcal{E}_K,$$

– S_N is the vector defined by

$$(S_N)_K = \frac{\mathrm{m}(K)}{\Delta t}\left(N_K^n + \frac{\alpha}{\lambda^2}\, N_K\right) + (Tb_N)_K \forall K\in\mathcal{T}$$

where

$$(Tb_N)_K = \begin{cases} 0 & \text{if } K\in\mathcal{T} \text{ is such that } \mathrm{m}(\partial K\cap\partial\Omega) = 0, \\ \displaystyle\sum_{\sigma\in\mathcal{E}_{K,ext}^D} \tau_\sigma\, B\left(D\Psi_{K,\sigma}\right)\, \tilde{N}_\sigma & \text{if } K\in\mathcal{T} \text{ is such that } \mathrm{m}(\partial K\cap\partial\Omega)\neq 0. \end{cases}$$

We can write in the same way the scheme (1.2.4) as $A_P\,\tilde{P} = S_P$. Since A_N and A_P have positive diagonal terms, nonpositive offdiagonal terms and are strictly diagonally dominant with respect to their columns, they are M-matrices. Then there exists a unique $(\tilde{N},\tilde{P})\in\mathbb{R}^\theta\times\mathbb{R}^\theta$ such that (1.2.3) and (1.2.4) are fulfilled, which gives that T_α^n is well-defined. Moreover, if $N,\ P\geq 0$, then $\tilde{N},\ \tilde{P}\geq 0$.
Now we will prove that T_α^n preserves the set

$$\mathcal{C} = \left\{(N,P)\in\mathbb{R}^\theta\times\mathbb{R}^\theta;\quad m\leq N_K\,P_K\leq M,\quad \forall K\in\mathcal{T}\right\}. \tag{1.2.5}$$

To this end, we first compute $A_N(\tilde{N}-\mathbf{M})$, where $\mathbf{M} = (M,...,M)^T\in\mathbb{R}^\theta$. Using the following property of the Bernoulli function

$$B(-x) - B(x) = x \quad \forall x\in\mathbb{R}, \tag{1.2.6}$$

we get that for all $K\in\mathcal{T}$,

$$\begin{aligned}\left(A_N(\tilde{N}-\mathbf{M})\right)_K =\ & \frac{\mathrm{m}(K)}{\Delta t}(N_K^n - M) + \frac{\mathrm{m}}{\Delta t}\,\frac{\alpha}{\lambda^2}(N_K - M) \\ & + \sum_{\sigma\in\mathcal{E}_{K,ext}^D} \tau_\sigma\left(B\left(D\Psi_{K,\sigma}\right) N_\sigma^{n+1} - B\left(-D\Psi_{K,\sigma}\right) M\right) \\ & - M\sum_{\sigma\in\mathcal{E}_{K,int}} \tau_\sigma\, D\Psi_{K,\sigma}.\end{aligned}$$

Thus since B is a nonnegative function and using the hypothesis (H3) we have

$$\begin{aligned}B\left(D\Psi_{K,\sigma}\right) N_\sigma^{n+1} - B\left(-D\Psi_{K,\sigma}\right) M =\ & B\left(D\Psi_{K,\sigma}\right)\left(N_\sigma^{n+1} - M\right) - D\Psi_{K,\sigma}\, M \\ \leq\ & -D\Psi_{K,\sigma}\, M \quad \forall\sigma\in\mathcal{E}_{K,ext}^D,\end{aligned}$$

it yields using the assumption of induction $N_K^n \leq M$ that

$$\Big(A_N(\tilde{N}-\mathbf{M})\Big)_K \leq \frac{\mathrm{m}(K)}{\Delta t}\,\frac{\alpha}{\lambda^2}\,(N_K-M) - M\sum_{\sigma\in\mathcal{E}_K}\tau_\sigma\, D\Psi_{K,\sigma}.$$

Finally using (1.2.2), we get

$$\Big(A_N(\tilde{N}-\mathbf{M})\Big)_K \leq \frac{\mathrm{m}(K)}{\lambda^2}\left(\frac{\alpha}{\Delta t}-M\right)(N_K-M) + M\,\frac{\mathrm{m}(K)}{\lambda^2}(P_K-M). \tag{1.2.7}$$

We can prove exactly in the same way that

$$\Big(A_N(\tilde{N}-\mathbf{m})\Big)_K \geq \frac{\mathrm{m}(K)}{\lambda^2}\left(\frac{\alpha}{\Delta t}-m\right)(N_K-m) + m\,\frac{\mathrm{m}(K)}{\lambda^2}(P_K-m). \tag{1.2.8}$$

Since $\alpha>0$ is an arbitrary constant, we can choose it such that $m\,\Delta t \leq M\,\Delta t \leq \alpha$, and then if $m\leq N_K$, $P_K\leq M$, we obtain from (1.2.7) that $(A_N(\tilde{N}-\mathbf{M}))_K\leq 0$ and from (1.2.8) that $(A_N(\tilde{N}-\mathbf{m}))_K\geq 0$, and then $m\leq \tilde{N}_K\leq M$ for all $K\in\mathcal{T}$ since A_N is a M-matrix. We can prove in the same way that $m\leq \tilde{P}_K\leq M$ for all $K\in\mathcal{T}$, and then $T_\alpha^n(\mathcal{C})\subset\mathcal{C}$, where $\mathcal{C}$ is defined by (1.2.5).
To conclude, T_α^n is a continuous application which stabilizes the set $\mathcal{C}$, then by the Brouwer fixed-point theorem, T_α^n has a fixed point (N^{n+1},P^{n+1}) in $\mathcal{C}$. Moreover, this fixed point with the corresponding Ψ is a solution to the scheme (1.1.13)–(1.1.14)–(1.1.15). This shows the existence of a solution to the scheme and the L^∞ estimate (1.2.1).

□

1.3 Properties of the numerical fluxes

In this section, we prove some useful properties of the Scharfetter-Gummel fluxes, which will be useful in all this part. We recall that the numerical flux $\mathcal{F}_{K,\sigma}^{n+1}$ is a numerical approximation of $\int_\sigma(-\nabla N+N\nabla\Psi)\cdot\mathbf{n}_{K,\sigma}$ on the interval $[t^n,t^{n+1})$. But, on the continuous level, we may rewrite $-\nabla N+N\nabla\Psi=-N\nabla(\log N-\Psi)$. Such an equality cannot be kept at the discrete level. However, we can give a lower and an upper bound of $\mathcal{F}_{K,\sigma}^{n+1}$ by terms of the form $-N_\sigma^{n+1}D(\log N-\Psi)_{K,\sigma}^{n+1}$, as shown in the following proposition:

Proposition 1.3.1. *For all $K\in\mathcal{T}$ and all $\sigma\in\mathcal{E}_{K,int}\cup\mathcal{E}_{K,ext}^D$, the flux $\mathcal{F}_{K,\sigma}^{n+1}$ defined by* (1.1.14) *satisfy the following inequalities:*

– *If $D(\log N-\Psi)_{K,\sigma}^{n+1}\geq 0$, then*

$$\begin{aligned}-\tau_\sigma\max(N_K^{n+1},N_{K,\sigma}^{n+1})\,D(\log N-\Psi)_{K,\sigma}^{n+1} \;&\leq\; \mathcal{F}_{K,\sigma}^{n+1}\\ &\leq\; -\tau_\sigma\min(N_K^{n+1},N_{K,\sigma}^{n+1})\,D(\log N-\Psi)_{K,\sigma}^{n+1}.\end{aligned} \tag{1.3.1}$$

– *If $D(\log N-\Psi)_{K,\sigma}^{n+1}\leq 0$, then*

$$\begin{aligned}-\tau_\sigma\min(N_K^{n+1},N_{K,\sigma}^{n+1})\,D(\log N-\Psi)_{K,\sigma}^{n+1} \;&\leq\; \mathcal{F}_{K,\sigma}^{n+1}\\ &\leq\; -\tau_\sigma\max(N_K^{n+1},N_{K,\sigma}^{n+1})\,D(\log N-\Psi)_{K,\sigma}^{n+1}.\end{aligned} \tag{1.3.2}$$

For all $K \in \mathcal{T}$ and all $\sigma \in \mathcal{E}_{K,int} \cup \mathcal{E}_{K,ext}^D$, the flux $\mathcal{G}_{K,\sigma}^{n+1}$ defined by (1.1.15) *satisfy the following inequalities:*

– *If $D(\log N + \Psi)_{K,\sigma}^{n+1} \geq 0$, then*

$$\begin{aligned} -\tau_\sigma \max(P_K^{n+1}, P_{K,\sigma}^{n+1})\, D(\log N + \Psi)_{K,\sigma}^{n+1} \;&\leq\; \mathcal{G}_{K,\sigma}^{n+1} \\ &\leq\; -\tau_\sigma \min(P_K^{n+1}, P_{K,\sigma}^{n+1})\, D(\log N + \Psi)_{K,\sigma}^{n+1}. \end{aligned}$$

– *If $D(\log N + \Psi)_{K,\sigma}^{n+1} \leq 0$, then*

$$\begin{aligned} -\tau_\sigma \min(P_K^{n+1}, P_{K,\sigma}^{n+1})\, D(\log N + \Psi)_{K,\sigma}^{n+1} \;&\leq\; \mathcal{G}_{K,\sigma}^{n+1} \\ &\leq\; -\tau_\sigma \max(P_K^{n+1}, P_{K,\sigma}^{n+1})\, D(\log N + \Psi)_{K,\sigma}^{n+1}. \end{aligned}$$

Proof. We focus on the flux $\mathcal{F}_{K,\sigma}^{n+1}$ defined by (1.1.14). Replacing Ψ by $-\Psi$ and N by P gives the result for $\mathcal{G}_{K,\sigma}^{n+1}$. Let $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_{K,int} \cup \mathcal{E}_{K,ext}^D$. Since the Bernoulli function satisfies (1.2.6), the flux $\mathcal{F}_{K,\sigma}^{n+1}$ defined by (1.1.14) can be either rewritten

$$\mathcal{F}_{K,\sigma}^{n+1} = \tau_\sigma \Big(D\Psi_{K,\sigma}^{n+1}\, N_K^{n+1} - B\left(D\Psi_{K,\sigma}^{n+1}\right) DN_{K,\sigma}^{n+1} \Big), \tag{1.3.3}$$

or

$$\mathcal{F}_{K,\sigma}^{n+1} = \tau_\sigma \Big(D\Psi_{K,\sigma}^{n+1}\, N_{K,\sigma}^{n+1} - B\left(-D\Psi_{K,\sigma}^{n+1}\right) DN_{K,\sigma}^{n+1} \Big). \tag{1.3.4}$$

It implies

$$\begin{aligned} \mathcal{F}_{K,\sigma}^{n+1} = \tau_\sigma \Big[D\Psi_{K,\sigma}^{n+1}\, N_K^{n+1} &- B\left(D(\log N)_{K,\sigma}^{n+1}\right) DN_{K,\sigma}^{n+1} \\ &+ \Big(B\left(D(\log N)_{K,\sigma}^{n+1}\right) - B\left(D\Psi_{K,\sigma}^{n+1}\right) \Big) DN_{K,\sigma}^{n+1} \Big], \end{aligned}$$

and

$$\begin{aligned} \mathcal{F}_{K,\sigma}^{n+1} = \tau_\sigma \Big[D\Psi_{K,\sigma}^{n+1}\, N_{K,\sigma}^{n+1} &- B\left(-D(\log N)_{K,\sigma}^{n+1}\right) DN_{K,\sigma}^{n+1} \\ &+ \Big(B\left(-D(\log N)_{K,\sigma}^{n+1}\right) - B\left(-D\Psi_{K,\sigma}^{n+1}\right) \Big) DN_{K,\sigma}^{n+1} \Big]. \end{aligned}$$

But the definition of the Bernoulli function (1.1.16) ensures that

$$B(\log y - \log x) = \frac{\log y - \log x}{y - x} x, \quad \forall x, y > 0.$$

Therefore, we get

$$\mathcal{F}_{K,\sigma}^{n+1} = \tau_\sigma \left[-D(\log N - \Psi)_{K,\sigma}^{n+1}\, N_K^{n+1} + \left(B\left(D(\log N)_{K,\sigma}^{n+1}\right) - B\left(D\Psi_{K,\sigma}^{n+1}\right) \right) DN_{K,\sigma}^{n+1} \right],$$

and

$$\mathcal{F}_{K,\sigma}^{n+1} = \tau_\sigma \left[-D(\log N - \Psi)_{K,\sigma}^{n+1} N_{K,\sigma}^{n+1} + \left(B\left(-D(\log N)_{K,\sigma}^{n+1} \right) - B\left(-D\Psi_{K,\sigma}^{n+1} \right) \right) DN_{K,\sigma}^{n+1} \right].$$

We may now use the fact that B is a nonincreasing function on $\mathbb{R}$. Assuming that the sign of $D(\log N - \Psi)_{K,\sigma}^{n+1}$ is known, the sign of $\left(B\left(D(\log N)_{K,\sigma}^{n+1} \right) - B\left(D\Psi_{K,\sigma}^{n+1} \right) \right)$ and $\left(B\left(-D(\log N)_{K,\sigma}^{n+1} \right) - B\left(-D\Psi_{K,\sigma}^{n+1} \right) \right)$ are also known (and opposite). Distinguishing the cases $DN_{K,\sigma}^{n+1} \geq 0$ ($N_K^{n+1} \leq N_{K,\sigma}^{n+1}$) and $DN_{K,\sigma}^{n+1} \leq 0$ ($N_K^{n+1} \geq N_{K,\sigma}^{n+1}$) yields the inequalities (1.3.1) and (1.3.2).

□

We now give a straightforward consequence of Proposition 1.3.1 as a corollary.

Corollary 1.3.1. *For all $K \in \mathcal{T}$ and all $\sigma \in \mathcal{E}_{K,int} \cup \mathcal{E}_{K,ext}^D$, the fluxes $\mathcal{F}_{K,\sigma}^{n+1}$ and $\mathcal{G}_{K,\sigma}^{n+1}$ defined respectively by* (1.1.14) *and* (1.1.15) *verify :*

$$\mathcal{F}_{K,\sigma}^{n+1}\ D(\log N - \Psi)_{K,\sigma}^{n+1}\ \leq -\tau_\sigma \min(N_K^{n+1}, N_{K,\sigma}^{n+1}) \left(D_\sigma (\log N - \Psi)^{n+1} \right)^2, \tag{1.3.5}$$

$$\mathcal{G}_{K,\sigma}^{n+1}\ D(\log P + \Psi)_{K,\sigma}^{n+1}\ \leq -\tau_\sigma \min(P_K^{n+1}, P_{K,\sigma}^{n+1}) \left(D_\sigma (\log P + \Psi)^{n+1} \right)^2. \tag{1.3.6}$$

Moreover, if $\min(N_K^{n+1}, N_{K,\sigma}^{n+1}) \geq 0$ *and* $\min(P_K^{n+1}, P_{K,\sigma}^{n+1}) \geq 0$, *we also have*

$$\left| \mathcal{F}_{K,\sigma}^{n+1} \right| \leq \tau_\sigma \max(N_K^{n+1}, N_{K,\sigma}^{n+1})\, D_\sigma \left(\log N - \Psi \right)^{n+1}, \tag{1.3.7}$$

$$\left| \mathcal{G}_{K,\sigma}^{n+1} \right| \leq \tau_\sigma \max(P_K^{n+1}, P_{K,\sigma}^{n+1})\, D_\sigma \left(\log P + \Psi \right)^{n+1}. \tag{1.3.8}$$

1.4 Long-time behavior of the Scharfetter-Gummel scheme

In this section we prove our main result, which is based on a discrete energy estimate with the control of the dissipation given in Proposition 1.4.1.

1.4.1 Notations and definitions

For $U = ((U_K)_{K\in\mathcal{T}}, (U_\sigma)_{\sigma\in\mathcal{E}_{ext}^D})$, we define the H^1-seminorm as follows:

$$|U|_{1,\Omega}^2 = \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma \left(D_\sigma U \right)^2.$$

Since the study of the large time behavior of the scheme (1.1.13)–(1.1.14)–(1.1.15) is based on an energy estimate with the control of its dissipation, let us introduce the discrete version of the deviation of the total energy from the thermal equilibrium defined by (1.1.8):

$$\begin{aligned} \mathcal{E}^n \ = \ & \sum_{K\in\mathcal{T}} \mathrm{m}(K) \left(H(N_K^n) - H(N_K^{eq}) - \log(N_K^{eq}) \left(N_K^n - N_K^{eq} \right) \right) \\ & + \sum_{K\in\mathcal{T}} \mathrm{m}(K) \left(H(P_K^n) - H(P_K^{eq}) - \log(P_K^{eq})(P_K^n - P_K^{eq}) \right) \\ & + \frac{\lambda^2}{2} |\Psi^n - \Psi^{eq}|_{1,\Omega}^2. \end{aligned}$$

Since $s \mapsto H(s) = \int_1^s \log(\tau) d\tau$ is defined and convex on $\mathbb{R}_+$, we have $\mathcal{E}^n \geq 0$ for all $n \geq 0$. We also introduce the discrete version of the energy dissipation defined by (1.1.9):

$$\begin{aligned}\mathcal{I}^n = & \sum_{\sigma \in \mathcal{E}_{int} \cup \mathcal{E}_{ext}^D} \tau_\sigma \min\left(N_K^n, N_{K,\sigma}^n\right) \left[D_\sigma \left(\log N - \Psi\right)^n\right]^2 \\ & + \sum_{\sigma \in \mathcal{E}_{int} \cup \mathcal{E}_{ext}^D} \tau_\sigma \min\left(P_K^n, P_{K,\sigma}^n\right) \left[D_\sigma \left(\log P + \Psi\right)^n\right]^2 .\end{aligned}$$

1.4.2 Energy estimate

The following proposition gives the control of energy and dissipation. This result is the keypoint to prove Theorem 1.1.1.

Proposition 1.4.1. *Under hypotheses (H1)–(H4), we have for all $n \geq 0$:*

$$0 \leq \mathcal{E}^{n+1} + \Delta t \, \mathcal{I}^{n+1} \leq \mathcal{E}^n. \tag{1.4.1}$$

Proof. Firstly, using the convexity of H and (1.1.12), we get

$$\begin{aligned}\mathcal{E}^{n+1} - \mathcal{E}^n \leq & \sum_{K \in \mathcal{T}} \mathrm{m}(K) \left(\log\left(N_K^{n+1}\right) - \alpha_N - \Psi_K^{eq}\right)\left(N_K^{n+1} - N_K^n\right) \\ & + \sum_{K \in \mathcal{T}} \mathrm{m}(K) \left(\log\left(P_K^{n+1}\right) - \alpha_P + \Psi_K^{eq}\right)\left(P_K^{n+1} - P_K^n\right) \\ & + \frac{\lambda^2}{2}\left|\Psi^{n+1} - \Psi^{eq}\right|_{1,\Omega}^2 - \frac{\lambda^2}{2}\left|\Psi^n - \Psi^{eq}\right|_{1,\Omega}^2,\end{aligned}$$

and then, by adding $\Psi_K^{n+1} - \Psi_K^{n+1}$ in the two first sums, we have

$$\mathcal{E}^{n+1} - \mathcal{E}^n \leq T_1 + T_2 + T_3,$$

where

$$\begin{aligned}T_1 &= \sum_{K \in \mathcal{T}} m(K) \left(\log\left(N_K^{n+1}\right) - \alpha_N - \Psi_K^{n+1}\right)\left(N_K^{n+1} - N_K^n\right), \\ T_2 &= \sum_{K \in \mathcal{T}} m(K) \left(\log\left(P_K^{n+1}\right) - \alpha_P + \Psi_K^{n+1}\right)\left(P_K^{n+1} - P_K^n\right), \\ T_3 &= \sum_{K \in \mathcal{T}} m(K) \left(\Psi_K^{n+1} - \Psi_K^{eq}\right)\left(N_K^{n+1} - N_K^n - P_K^{n+1} + P_K^n\right) \\ & \quad + \frac{\lambda^2}{2}\left|\Psi^{n+1} - \Psi^{eq}\right|_{1,\Omega}^2 - \frac{\lambda^2}{2}\left|\Psi^n - \Psi^{eq}\right|_{1,\Omega}^2.\end{aligned}$$

The scheme (1.1.13) on Ψ and a discrete integration by parts (using that $\Psi_\sigma^{n+1} = \Psi_\sigma^{eq}$ for all $\sigma \in \mathcal{E}_{ext}^D$) yield:

$$\begin{aligned} T_3 =& -\lambda^2 \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\, D\left(\Psi^{n+1}-\Psi^{eq}\right)_{K,\sigma} D\left(\Psi^{n}-\Psi^{eq}\right)_{K,\sigma} \\ &+ \frac{\lambda^2}{2} \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma \left(\left(D_\sigma\left(\Psi^{n+1}-\Psi^{eq}\right)\right)^2 - \left(D_\sigma\left(\Psi^{n}-\Psi^{eq}\right)\right)^2\right) \\ =& -\frac{\lambda^2}{2}\left|\Psi^{n+1}-\Psi^{eq}\right|_{1,\Omega}^2 \le 0. \end{aligned}$$

Then we use the scheme on N and a discrete integration by parts (using that $\log(N_\sigma^{n+1}) - \alpha_N - V_\sigma^{n+1} = 0$ for all $\sigma \in \mathcal{E}_{ext}^D$ by hypothesis (H4)) to obtain:

$$T_1 = \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\, D\left(\log N-\Psi\right)_{K,\sigma}^{n+1} \mathcal{F}_{K,\sigma}^{n+1}.$$

Thus using (1.3.5), we get

$$T_1 \le -\Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\, \min\left(N_K^{n+1}, N_{K,\sigma}^{n+1}\right)\left(D_\sigma\left(\log N-\Psi\right)^{n+1}\right)^2,$$

and we obtain in the same way a similar estimate for T_2 by using (1.3.6):

$$T_2 \le -\Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\, \min\left(P_K^{n+1}, P_{K,\sigma}^{n+1}\right)\left(D_\sigma\left(\log P+\Psi\right)^{n+1}\right)^2.$$

To sum up, we have

$$\mathcal{E}^{n+1} - \mathcal{E}^n \le T_1 + T_2 \le -\Delta t\, \mathcal{I}^{n+1},$$

which completes the proof. □

1.4.3 Proof of Theorem 1.1.1

Now we are ready to achieve the proof of Theorem 1.1.1. We follow the same ideas as in [51]. We first obtain by summing (1.4.1) over $k = 0, ..., n$ that for all $n \ge 0$,

$$0 \le \mathcal{E}^{n+1} + \sum_{k=0}^{n} \Delta t\, \mathcal{I}^{k+1} \le \mathcal{E}^0.$$

Then the series $\sum_k \mathcal{I}^{k+1}$ is bounded, and since $\mathcal{I}^{k+1} \ge 0$ for all k, thus

$$\mathcal{I}^k \to 0 \text{ as } k \to +\infty. \tag{1.4.2}$$

Applying Young and Cauchy-Schwarz inequalities, we get for any $\eta > 0$:

$$\begin{aligned}\frac{\eta}{2}\sum_{K\in\mathcal{T}}\mathrm{m}(K)\left(N_K^{n+1}-N_K^{eq}\right)^2+\frac{1}{2\,\eta}\sum_{K\in\mathcal{T}}\mathrm{m}(K)\left(\log(N_K^{n+1})-\alpha_N-\Psi_K^{n+1}\right)^2\\ \geq\sum_{K\in\mathcal{T}}\mathrm{m}(K)\left(N_K^{n+1}-N_K^{eq}\right)\left(\log(N_K^{n+1})-\alpha_N-\Psi_K^{n+1}\right)\\ =\sum_{K\in\mathcal{T}}\mathrm{m}(K)\left(N_K^{n+1}-N_K^{eq}\right)\left(\log(N_K^{n+1})-\log(N_K^{eq})+\Psi_K^{eq}-\Psi_K^{n+1}\right)\end{aligned}$$

by using hypothesis (H4). Moreover, since there exists $c > 0$ such that for all x, $y \in [m, M]$,

$$(x-y)\left(\log(x)-\log(y)\right)\geq c(x-y)^2,$$

we get

$$\begin{aligned}\frac{\eta}{2}\sum_{K\in\mathcal{T}}\mathrm{m}(K)\left(N_K^{n+1}-N_K^{eq}\right)^2+\frac{1}{2\,\eta}\sum_{K\in\mathcal{T}}\mathrm{m}(K)\left(\log(N_K^{n+1})-\alpha_N-\Psi_K^{n+1}\right)^2\geq\\ c\sum_{K\in\mathcal{T}}\mathrm{m}(K)\left(N_K^{n+1}-N_K^{eq}\right)^2+\sum_{K\in\mathcal{T}}\mathrm{m}(K)\left(N_K^{n+1}-N_K^{eq}\right)\left(\Psi_K^{eq}-\Psi_K^{n+1}\right),\end{aligned}$$

and in the same way

$$\begin{aligned}\frac{\eta}{2}\sum_{K\in\mathcal{T}}\mathrm{m}(K)\left(P_K^{n+1}-P_K^{eq}\right)^2+\frac{1}{2\,\eta}\sum_{K\in\mathcal{T}}\mathrm{m}(K)\left(\log(P_K^{n+1})-\alpha_P+\Psi_K^{n+1}\right)^2\geq\\ c\sum_{K\in\mathcal{T}}\mathrm{m}(K)\left(P_K^{n+1}-P_K^{eq}\right)^2-\sum_{K\in\mathcal{T}}\mathrm{m}(K)\left(P_K^{n+1}-P_K^{eq}\right)\left(\Psi_K^{eq}-\Psi_K^{n+1}\right).\end{aligned}$$

Thus, adding the two latter inequalities, using the schemes (1.1.13) on Ψ^{n+1} and (1.1.12) on Ψ^{eq}, and a discrete integration by parts, we obtain for $\eta < 2\,c$:

$$\begin{aligned}\left(c-\frac{\eta}{2}\right)\sum_{K\in\mathcal{T}}\mathrm{m}(K)\left[\left(N_K^{n+1}-N_K^{eq}\right)^2+\left(P_K^{n+1}-P_K^{eq}\right)^2\right]+\lambda^2\left|\Psi^{n+1}-\Psi^{eq}\right|_{1,\Omega}^2\leq\\ \frac{1}{2\,\eta}\sum_{K\in\mathcal{T}}\mathrm{m}(K)\left[\left(\log(N_K^{n+1})-\alpha_N-\Psi_K^{n+1}\right)^2+\left(\log(P_K^{n+1})-\alpha_P+\Psi_K^{n+1}\right)^2\right].\end{aligned}$$

Now since $\log(N_\sigma^{n+1})-\alpha_N-\Psi_\sigma^{n+1}=0$ and $\log(P_\sigma^{n+1})-\alpha_P+\Psi_\sigma^{n+1}=0$ for all $\sigma\subset\Gamma^D$ by hypothesis (H4) and since the mesh satisfies the regularity constraint (1.1.10), we can apply a discrete Poincaré inequality (see Theorem 5.4.2 in Chapter 5) and obtain that there exists a constant $C > 0$ only depending on Ω and ξ such that

$$\begin{aligned}\left(c-\frac{\eta}{2}\right)\sum_{K\in\mathcal{T}}\mathrm{m}(K)\left[\left(N_K^{n+1}-N_K^{eq}\right)^2+\left(P_K^{n+1}-P_K^{eq}\right)^2\right]+\lambda^2\left|\Psi^{n+1}-\Psi^{eq}\right|_{1,\Omega}^2\leq\\ \frac{C}{2\,\eta}\left(\left|\log(N^{n+1})-\Psi^{n+1}\right|_{1,\Omega}^2+\left|\log(P^{n+1})+\Psi^{n+1}\right|_{1,\Omega}^2\right).\end{aligned}$$

Finally using the uniform lower bound (1.2.1) we get

$$\left(c-\frac{\eta}{2}\right)\sum_{K\in\mathcal{T}}\mathrm{m}(K)\left[\left(N_K^{n+1}-N_K^{eq}\right)^2+\left(P_K^{n+1}-P_K^{eq}\right)^2\right]$$
$$+\lambda^2\left|\Psi^{n+1}-\Psi^{eq}\right|_{1,\Omega}^2\leq\frac{C}{2\,\eta\,m}\mathcal{I}^{n+1}.$$

Therefore, passing to the limit $n\to\infty$ and using (1.4.2), we finally get the result:

$$N_K^n\to N_K^{eq},\quad P_K^n\to P_K^{eq},\quad \Psi_K^n\to\Psi_K^{eq}\quad\text{as } n\to\infty.$$

1.5 Numerical experiments

In this section, we give numerical results in one and two space dimensions, obtained on the one hand with the Scharfetter-Gummel scheme (1.1.13)–(1.1.14)–(1.1.15) and on the other hand with the scheme studied by C. Chainais-Hillairet, J.-G. Liu and Y.-J. Peng in [52], where the diffusion terms are discretized classically with two-points fluxes and the convection terms are discretized with upwind fluxes. Moreover, we compute an approximation of the thermal equilibrium (1.1.6) with the scheme proposed and studied in [51].

Test case 1. We first present a test case for a geometry corresponding to a PN-junction in 1D. The domain is $\Omega=(0,1)$. The doping profile C is given by

$$C(x)=\begin{cases}-1 & \text{if } x\in[0,1/2),\\ +1 & \text{elsewhere.}\end{cases}$$

The rescaled Debye length λ is taken equal to 10^{-2}. Moreover, Dirichlet boundary conditions are prescribed

$$N^D(0)=P^D(1)=0.01,\quad P^D(0)=N^D(1)=1,$$

and the potential Ψ^D at the boundary is such that the compatibility condition (1.1.5) is fulfilled:

$$\Psi^D(\gamma)=\frac{\log(N^D(\gamma))-\log(P^D(\gamma))}{2},\quad \gamma=\{0,1\}.$$

We run computations with a time step $\Delta t=10^{-3}$ on a mesh made of 100 cells. In Figure 1.1 we compare the relative energy $\mathcal{E}^n$ and its dissipation $\mathcal{I}^n$ obtained with the the Scharfetter–Gummel scheme (1.1.13)–(1.1.14)–(1.1.15) and with the scheme studied in [52]. With the Scharfetter-Gummel scheme, we observe that $\mathcal{E}^n$ and $\mathcal{I}^n$ converge to zero when $n\to\infty$, which is in keeping with Theorem 1.1.1 even if the doping profile C is not zero. Moreover, from the numerical results, it seems that the convergence is exponential in time, as proved at the continuous level by H. Gajewski and K. Gärtner in [92]. On the contrary, the upwind scheme, which does not preserve thermal equilibrium, is not very satisfying to reflect the long time behavior of the solution.

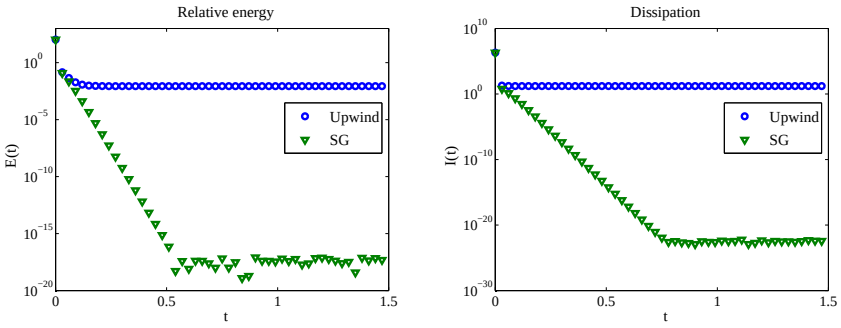


Figure 1.1: Test case 1: evolution of the relative energy $\mathcal{E}^n$ and its dissipation $\mathcal{I}^n$ in log-scale.

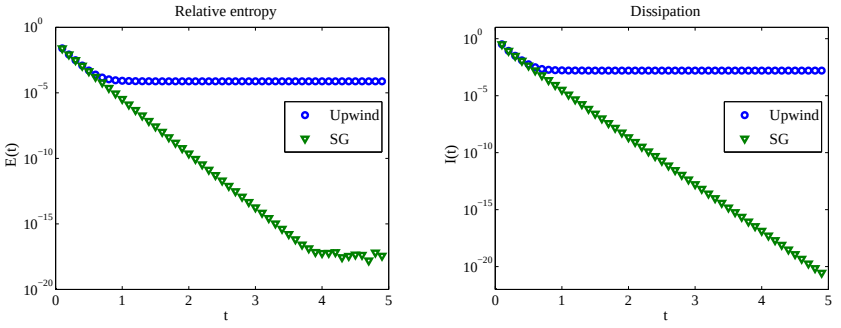


Figure 1.2: Test case 2: evolution of the relative energy $\mathcal{E}^n$ and its dissipation $\mathcal{I}^n$ in log-scale.

Test case 2. We now consider a 2D test case picked in the paper of C. Chainais-Hillairet and F. Filbet [51]. The domain Ω is the square $(0,1)^2$. The doping profile is piecewise constant, equal to +1 in the N-region and -1 in the P-region. The rescaled Debye length is taken equal to $\lambda = 1$. The Dirichlet boundary conditions are

$$N^D = 0.1, \quad P^D = 0.9, \quad \Psi^D = \frac{\log(N^D) - \log(P^D)}{2} \quad \text{on } \{y = 1, 0 \leq x \leq 0.25\},$$
$$N^D = 0.9, \quad P^D = 0.1, \quad \Psi^D = \frac{\log(N^D) - \log(P^D)}{2} \quad \text{on } \{y = 0\}.$$

Elsewhere we put homogeneous Neumann boundary conditions. We compute the numerical approximations of the thermal equilibrium and of the transient drift-diffusion system on a mesh made of 896 triangles, with time step $\Delta t = 10^{-2}$. Figure 1.2 represents the evolution of the relative energy and its dissipation obtained with the Scharfetter-Gummel scheme and the scheme proposed in [52] in log scale. On the one hand, it highlights an exponential behavior with respect to the time for the Scharfetter-Gummel scheme. On the other hand, we still observe a phenomenon of saturation with the other scheme.

Chapitre 2

Stabilité du schéma de Scharfetter-Gummel à la limite quasi-neutre *

Dans ce chapitre, nous considérons toujours le schéma de Scharfetter-Gummel implicite pour le modèle de dérive-diffusion linéaire pour les semi-conducteurs. Cependant, nous nous intéressons cette fois-ci à la préservation d'une autre asymptotique : la limite quasi-neutre $\lambda \to 0$. Nous prouvons des estimations a priori indépendantes de la longueur de Debye λ, ce qui permet de montrer la convergence du schéma pour tout $\lambda > 0$. Le point clé est à nouveau une estimation d'entropie–dissipation d'entropie discrète. Ce travail constitue la première étape pour montrer que le schéma de Scharfetter-Gummel préserve l'asymptotique quasi-neutre.

∗. Ce chapitre est un travail effectué en collaboration avec C. Chainais-Hillairet et M.-H. Vignal. L'article correspondant, *Convergence of a fully implicit scheme for the drift-diffusion system. Stability at the quasineutral limit* [18], est actuellement en cours d'écriture.

Contents

2.1 Introduction

2.1.1 Aim of the chapter

In the modeling of plasmas or semiconductor devices, there is a hierarchy of different models: kinetic models and quasi hydrodynamic models, ranging from Euler-Poisson system to drift-diffusion systems [120, 137]. In each of these models scaled parameters are involved, like the mass of electrons, the relaxation time or the Debye length. An active field of research consists in designing numerical schemes for these physical models which are valid for all range of scaled parameters, and especially when these parameters may tend to 0.

In this chapter, we will focus on the linear drift-diffusion system. It is a coupled system of parabolic and elliptic equations involving only one dimensionless parameter: λ, the rescaled Debye length which is given by the ratio of the Debye length to the size of the domain. The Debye length measures the typical scale of electric interactions in the semiconductor. We are interested in the so-called quasi-neutral regime. This regime occurs when the parameter λ tends to zero. We will establish that the fully implicit in time Scharfetter-Gummel finite volume scheme for the drift-diffusion system converges for all $\lambda > 0$.

Many different numerical methods have been already developed for the approximation of the drift-diffusion system; see for instance the mixed exponential fitting schemes proposed in [34] and extended in [118, 122] to the case of nonlinear diffusion. The convergence of some finite volume schemes has been proved by C. Chainais-Hillairet, J.-G. Liu and Y.-J. Peng in [52, 54]. But, up to our knowledge, all the schemes are studied in the case $\lambda = 1$ and the behavior when λ tends to 0 has not yet been studied.

In this work, we are interested in proving that the implicit Scharfetter-Gummel scheme converges for any values of $\lambda > 0$. To this end, we want to follow the different steps of

the proof in [54]: as it is classical in the finite volume framework (see [85]), we first need a priori estimates on the approximate solutions (L^∞ and discrete $L^2(0,T,H^1)$ estimates) which yields compactness of the sequence of approximate solutions and then we may pass to the limit in the scheme. However, the crucial point in our work is to establish that all the a priori estimates do not depend of λ and therefore the strategy used in [54] to get them does not directly apply. In order to get estimates which are independent of λ, we will adapt the entropy method proposed in [121, 94, 95] at the discrete level. The choice of the Scharfetter-Gummel fluxes for the discretization of the convection-diffusion fluxes will be essential at this step.

2.1.2 The continuous problem

We recall here the continuous framework. Let $\Omega \subset \mathbb{R}^d$ ($d \geq 1$) be an open bounded domain describing the geometry of a semiconductor device and $T > 0$. The unknowns of the linear drift-diffusion system are the density of electrons and holes, N and P, and the electrostatic potential Ψ. It writes for all $(x,t) \in \Omega \times [0,T]$:

$$\left\{ \begin{array}{l} \partial_t N + \mathrm{div}(-\nabla N + N\nabla\Psi) = 0, \\ \\ \partial_t P + \mathrm{div}(-\nabla P - P\nabla\Psi) = 0, \\ \\ -\lambda^2 \Delta\Psi = P - N + C, \end{array} \right. \tag{2.1.1}$$

where $\lambda \geq 0$ is given and where C is the doping profile characterizing the device and depending only on x. We consider mixed boundary conditions as in [121]: Dirichlet boundary conditions on the ohmic contacts and homogeneous boundary conditions on the insulated boundary segments. It means that the boundary $\partial\Omega$ is splitted into $\partial\Omega = \Gamma^D \cup \Gamma^N$ with $\Gamma^D \cap \Gamma^N = \emptyset$ and that the boundary conditions write:

$$N(\gamma,t) = N^D(\gamma),\ P(\gamma,t) = P^D(\gamma), \Psi(\gamma,t) = \Psi^D(\gamma),\ (\gamma,t) \in \Gamma^D \times [0,T], \tag{2.1.2}$$

$$\nabla N \cdot \nu = \nabla P \cdot \nu = \nabla\Psi \cdot \nu = 0 \text{ on } \Gamma^N \times [0,T]. \tag{2.1.3}$$

The system (2.1.1) is also supplemented with initial conditions N_0, P_0:

$$N(x,0) = N_0(x), \quad P(x,0) = P_0(x), \quad x \in \Omega. \tag{2.1.4}$$

In the sequel, we need the following assumptions on the data:

Hypotheses H.1. *We assume that the boundary conditions N^D, P^D and Ψ^D are the traces of some functions defined on the whole domain Ω, still denoted by N^D, P^D and Ψ^D. We also assume that*

$$N_0, P_0 \in L^\infty(\Omega), \tag{2.1.5}$$

$$N^D, P^D \in L^\infty \cap H^1(\Omega), \Psi^D \in H^1(\Omega), \tag{2.1.6}$$

$$C \in L^\infty(\Omega), \tag{2.1.7}$$

$$\exists m > 0, M > 0 \text{ such that } m \leq N_0, P_0, N^D, P^D \leq M \text{ a.e.} \tag{2.1.8}$$

We define the notion of weak solution to the drift-diffusion system (2.1.1)–(2.1.4).

Definition 2.1. Under assumptions H.1, the triplet (N, P, Ψ) is a solution to the problem (2.1.1)–(2.1.4) if it satisfies: $N,\ P \in L^\infty(\Omega \times (0,T))$, $N - N^D,\ P - P^D,\ \Psi - \Psi^D \in L^2(0,T;V)$, with

$$V = \Big\{ v \in H^1(\Omega)\ ;\ v = 0 \text{ on } \Gamma^D \Big\},$$

and, for all test functions $\phi \in \mathcal{D}(\Omega \times [0,T))$ and $\eta \in \mathcal{D}(\Omega \times (0,T))$,

$$\int_0^T \int_\Omega (N\,\partial_t\phi - \nabla N \cdot \nabla\phi + N\,\nabla\Psi \cdot \nabla\phi)\,dx\,dt + \int_\Omega N_0(x)\,\phi(x,0)\,dx = 0, \tag{2.1.9}$$

$$\int_0^T \int_\Omega (P\,\partial_t\phi - \nabla P \cdot \nabla\phi - P\,\nabla\Psi \cdot \nabla\phi)\,dx\,dt + \int_\Omega P_0(x)\,\phi(x,0)\,dx = 0, \tag{2.1.10}$$

$$\lambda^2 \int_0^T \int_\Omega \nabla\Psi \cdot \nabla\eta\,dx\,dt = \int_0^T \int_\Omega (P - N + C)\,\eta\,dx\,dt. \tag{2.1.11}$$

In [121], A. Jüngel and Y.-J. Peng perform rigorously the quasi-neutral limit for the drift-diffusion system with a zero doping profile and mixed Dirichlet and homogeneous Neumann boundary conditions. The same kind of result is established for the drift-diffusion system with homogeneous Neumann boundary conditions by I. Gasser in [94] for a zero doping profile and by I. Gasser, C. D. Levermore, P. Markowich, C. Schmeiser in [95] for a regular doping profile. In all these papers, the rigorous proof of the quasi-neutral limit is based on an entropy method.
In this case, the entropy functional is defined (see [121]) by

$$\begin{aligned}\mathbb{E}(t) = \int_\Omega \Big(& H(N) - H(N^D) - \log(N^D)(N - N^D) \\ & + H(P) - H(P^D) - \log(P^D)(P - P^D) \\ & + \frac{\lambda^2}{2}|\nabla\Psi - \nabla\Psi^D|^2 \Big)\,dx,\end{aligned}$$

with $H(x) = \displaystyle\int_1^x \log(x)\ dx = x\log x - x + 1$, and the entropy production is defined by

$$\mathbb{I}(t) = \int_\Omega \Big(|\nabla(\log N - \Psi)|^2 + |\nabla(\log P + \Psi)|^2 \Big)\,dx\,dt.$$

The entropy inequality proved in [121]

$$\mathbb{E}(t) - \mathbb{E}(0) + \int_0^t \mathbb{I}(s)\,ds \le 0 \tag{2.1.12}$$

is crucial for the study of the quasi-neutral limit because it provides a priori estimates which are essential when performing rigorously the limit.

2.1.3 Presentation of the scheme

We will consider the same notations as in Chapter 1. A mesh of Ω is given by $\mathcal{T}$, a family of open control volumes, $\mathcal{E}$, a family of edges and $\mathcal{P} = (x_K)_{K\in\mathcal{T}}$ a family of points. As it is classical in the finite volume discretization of elliptic equations with a two-points flux approximation, we assume that the mesh is admissible in the sense of [85, Definition 9.1]. It implies that the straight line between two neighboring centers of cell (x_K, x_L) is orthogonal to the edge $\sigma = K|L$. The set of edges will be split into $\mathcal{E} = \mathcal{E}_{int} \cup \mathcal{E}_{ext}$ and for the exterior edges, we distinguish the edges included in Γ^D from the edges included in Γ^N: $\mathcal{E}_{ext} = \mathcal{E}_{ext}^D \cup \mathcal{E}_{ext}^N$. For a given control volume $K \in \mathcal{T}$, we define $\mathcal{E}_K$ the set of its edges, which is also split into $\mathcal{E}_K = \mathcal{E}_{K,int} \cup \mathcal{E}_{K,ext}^D \cup \mathcal{E}_{K,ext}^N$. For all edge $\sigma \in \mathcal{E}$, we define $\mathrm{d}_\sigma = \mathrm{d}(x_K, x_L)$ if $\sigma = K|L \in \mathcal{E}_{int}$ and $\mathrm{d}_\sigma = \mathrm{d}(x_K, \sigma)$ if $\sigma \in \mathcal{E}_{K,ext}$. Then, the transmissibility coefficient is defined by $\tau_\sigma = \mathrm{m}(\sigma)/\mathrm{d}_\sigma$, for all $\sigma \in \mathcal{E}$. Let $\Delta t > 0$ be the time step. We set $N_T = E(T/\Delta t)$ and $t^n = n\Delta t$ for all $0 \le n \le N_T$. We define the size of the time-space discretization by $\delta = \max(\Delta t, \mathrm{size}(\mathcal{T}))$, where $\mathrm{size}(\mathcal{T}) = \max_{K\in\mathcal{T}} [\mathrm{diam}(K)]$.

For almost all $t \in (0,T)$, the continuous solution (N, P, Ψ) is in $H^1(\Omega)$ and its trace is known on Γ^D. Then, for each time step we approximate $X = N$, P or Ψ by a constant function given on each control volume $K \in \mathcal{T}$ and each boundary edge $\sigma \in \mathcal{E}_{ext}^D \subset \Gamma^D$ by:

$$X_{\mathcal{T}}^{n+1} = ((X_K^{n+1})_{K\in\mathcal{T}}, (X_\sigma^{n+1})_{\sigma\in\mathcal{E}_{ext}^D}).$$

For $n = 0$, we discretize the initial conditions:

$$N_K^0 = \frac{1}{\mathrm{m}(K)} \int_K N_0(x)\,dx, \quad P_K^0 = \frac{1}{\mathrm{m}(K)} \int_K P_0(x)\,dx, \quad \forall K \in \mathcal{T}. \tag{2.1.13}$$

Similarly, the boundary conditions give:

$$N_\sigma^{n+1} = N_\sigma^D, \quad P_\sigma^{n+1} = P_\sigma^D, \quad \Psi_\sigma^{n+1} = \Psi_\sigma^D, \quad \forall \sigma \in \mathcal{E}_{ext}^D, \quad \forall n \ge 0, \tag{2.1.14}$$

where for $X = N$, P, or Ψ,

$$X_\sigma^D = \frac{1}{\mathrm{m}(\sigma)} \int_\sigma X^D(\gamma)\,d\gamma, \quad \forall \sigma \in \mathcal{E}_{ext}^D. \tag{2.1.15}$$

Moreover we assume Hypotheses H.1, then the boundary conditions N^D, P^D, Ψ^D are extended on the whole domain Ω, and we define for $X = N$, P, or Ψ,

$$X_K^D = \frac{1}{\mathrm{m}(K)} \int_K X^D(x)\,dx, \quad \forall K \in \mathcal{T}.$$

We consider the same scheme as in Chapter 1:

$$\begin{cases} \mathrm{m}(K)\dfrac{N_K^{n+1} - N_K^n}{\Delta t} + \displaystyle\sum_{\sigma\in\mathcal{E}_K} \mathcal{F}_{K,\sigma}^{n+1} = 0, & \forall K \in \mathcal{T}, \forall n \ge 0, \\ \mathrm{m}(K)\dfrac{P_K^{n+1} - P_K^n}{\Delta t} + \displaystyle\sum_{\sigma\in\mathcal{E}_K} \mathcal{G}_{K,\sigma}^{n+1} = 0, & \forall K \in \mathcal{T}, \forall n \ge 0, \\ \lambda^2 \displaystyle\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma D\Psi_{K,\sigma}^n = \mathrm{m}(K)\left(N_K^n - P_K^n - C_K\right), & \forall K \in \mathcal{T}, \forall n \ge 0, \end{cases} \tag{2.1.16}$$

where for any discrete quantity $X = ((X_K)_{K\in\mathcal{T}}, (X^D_\sigma)_{\sigma\in\mathcal{E}^D_{ext}})$, for all $K \in \mathcal{T}$ and all $\sigma \in \mathcal{E}_{K,int} \cup \mathcal{E}^D_{K,ext}$, we set

$$X_{K,\sigma} = \begin{cases} X_L, & \text{if } \sigma = K|L \in \mathcal{E}_{K,int}, \\ X_\sigma, & \text{if } \sigma \in \mathcal{E}^D_{K,ext}, \end{cases}$$

and

$$DX_{K,\sigma} = \begin{cases} X_{K,\sigma} - X_K, & \text{if} \quad \sigma \in \mathcal{E}_{K,int} \cup \mathcal{E}^D_{K,ext}, \\ 0, & \text{if} \quad \sigma \in \mathcal{E}^N_{K,ext}. \end{cases}$$

Finally we also define for $\sigma \in \mathcal{E}_K$:

$$D_\sigma X = |DX_{K,\sigma}|\,.$$

The quantities $\mathcal{F}^{n+1}_{K,\sigma}$ and $\mathcal{G}^{n+1}_{K,\sigma}$ are numerical approximations of $\int_\sigma(-\nabla N + N\nabla\Psi)\cdot\mathbf{n}_{K,\sigma}$ and $\int_\sigma(-\nabla P - P\nabla\Psi)\cdot\mathbf{n}_{K,\sigma}$ on the interval $[t^n, t^{n+1})$. The diffusive part and the convective part of the fluxes are discretized simultaneously by using the Scharfetter-Gummel fluxes defined by:

(2.1.17)
$$\mathcal{F}^{n+1}_{K,\sigma} = \begin{cases} \tau_\sigma \left(B\left(-D\Psi^{n+1}_{K,\sigma}\right) N^{n+1}_K - B\left(D\Psi^{n+1}_{K,\sigma}\right) N^{n+1}_{K,\sigma} \right) & \text{if} \quad \sigma \in \mathcal{E}_{K,int} \cup \mathcal{E}^D_{K,ext}, \\ 0 & \text{if} \quad \sigma \in \mathcal{E}^N_{K,ext}, \end{cases}$$

(2.1.18)
$$\mathcal{G}^{n+1}_{K,\sigma} = \begin{cases} \tau_\sigma \left(B\left(D\Psi^{n+1}_{K,\sigma}\right) P^{n+1}_K - B\left(-D\Psi^{n+1}_{K,\sigma}\right) P^{n+1}_{K,\sigma} \right) & \text{if} \quad \sigma \in \mathcal{E}_{K,int} \cup \mathcal{E}^D_{K,ext}, \\ 0 & \text{if} \quad \sigma \in \mathcal{E}^N_{K,ext}, \end{cases}$$

where B is the Bernoulli function defined by

$$B(0) = 1 \text{ and } B(x) = \frac{x}{\exp(x) - 1} \ \forall x \neq 0. \tag{2.1.19}$$

The scheme (2.1.13)–(2.1.19) is fully implicit in time: the numerical solution at each time step is defined as a solution of the nonlinear system of equations (2.1.16)–(2.1.18). When choosing $D\Psi^n_{K,\sigma}$ instead of $D\Psi^{n+1}_{K,\sigma}$ in the definition of the fluxes (2.1.17)–(2.1.18), we would get a decoupled scheme whose solution is obtained by solving successively three linear systems of equations for N, P and Ψ. However, this other choice of time discretization used in [52, 54] induces a stability condition of the form $\Delta t \leq C\lambda^2$ (see [14]) and therefore cannot be used in practice for small values of λ.
We recall here the result of existence and L^∞ stability obtained in Theorem 1.2.1, Chapter 1:

Theorem 2.1.1. *Let us assume* (2.1.5)–(2.1.6)–(2.1.8) *and that the doping profile C vanishes on the whole domain. Then there exists a solution* $\{(N^n_K, P^n_K, \Psi^n_K),\ K \in \mathcal{T},\ 0 \leq n \leq N_T\}$ *to the scheme* (2.1.13)–(2.1.19)*, satisfying the following L^∞ estimate:*

$$0 < m \leq N^n_K,\ P^n_K \leq M, \quad \forall K \in \mathcal{T}, \quad \forall n \geq 0. \tag{2.1.20}$$

Then an approximate solution $(N_\delta, P_\delta, \Psi_\delta)$ to the problem (2.1.1)–(2.1.4) associated to the discretization $\mathcal{D}$ of size δ is defined as a piecewise constant function by:

$$N_\delta(x,t) = N_K^{n+1}, \quad P_\delta(x,t) = P_K^{n+1}, \quad \Psi_\delta(x,t) = \Psi_K^{n+1}, \quad \forall (x,t) \in K \times [t^n, t^{n+1}[,$$

where $\{(N_K^n, P_K^n, \Psi_K^n),\ K \in \mathcal{T},\ 0 \le n \le N_T\}$ is a solution to the scheme (2.1.13)–(2.1.19). We also define an approximation of the gradients of the densities N and P, and of the electric potential Ψ as in [52]. Therefore, we define a dual mesh; for $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$, we define $T_{K,\sigma}$ as follows:

- if $\sigma = K|L \in \mathcal{E}_{K,int}$, then $T_{K,\sigma}$ is the cell whose vertices are x_K, x_L and those of $\sigma = K|L$,
- if $\sigma \in \mathcal{E}_{K,ext}$, then $T_{K,\sigma}$ is the cell whose vertices are x_K and those of σ.

Then $(T_{K,\sigma})_{\sigma \in \mathcal{E}_K,\, K \in \mathcal{T}}$ defines a partition of Ω. For $X = ((X_K^n)_{K\in\mathcal{T}, n\ge 0}, (X_\sigma^D)_{\sigma \in \mathcal{E}_{ext}^D})$, the approximation dX_δ is a piecewise constant function defined in $\Omega \times (0,T)$ by:

$$dX_\delta(x,t) = \frac{\mathrm{m}(\sigma)}{\mathrm{m}(T_{K,\sigma})}\, DX_{K,\sigma}^{n+1} \quad \text{if } (x,t) \in T_{K,\sigma} \times [t^n, t^{n+1}[.$$

2.1.4 Main result and outline of the chapter

We now briefly outline the contents of this chapter. In the next section we will prove some discrete a priori estimates independent of the rescaled Debye length $\lambda > 0$, based on a discrete entropy estimate. Then Section 2.3 is devoted to the proof of the convergence of the scheme (2.1.13)–(2.1.19) for all $\lambda > 0$, which is our main result, summarized in the following theorem:

Theorem 2.1.2. *Let us assume* (2.1.5)–(2.1.6)–(2.1.8)*, and the doping profile vanishes on the whole domain:* $C = 0$*. We further assume that the boundary and initial conditions satisfy the quasi-neutrality assumption*

$$N^D = P^D \text{ in } \Omega, \tag{2.1.21}$$

$$N^0 = P^0 \text{ in } \Omega. \tag{2.1.22}$$

Then there exists $N,\ P \in L^2(0,T;H^1(\Omega))$ *and* $\Psi \in \left(L^2(0,T;H^1(\Omega))\right)^d$ *such that the solution* $(N_\delta, P_\delta, \Psi_\delta)$ *given by the scheme* (2.1.13)–(2.1.19) *satisfies, up to subsequences,*

$$N_\delta \to N \text{ and } P_\delta \to P \text{ in } L^2(\Omega\times(0,T)) \text{ strongly}, \quad \text{as } \delta \to 0,$$

$$\Psi_\delta \rightharpoonup \Psi \text{ in } \left(L^2(\Omega\times(0,T))\right)^d \text{ weakly}, \quad \text{as } \delta \to 0,$$

$$dN_\delta \rightharpoonup \nabla N, \quad dP_\delta \rightharpoonup \nabla P, \quad d\Psi_\delta \rightharpoonup \nabla \Psi \text{ in } \left(L^2(\Omega\times(0,T))\right)^d \text{ weakly}, \quad \text{as } \delta \to 0.$$

Moreover, (N, P, Ψ) *is a weak solution to the drift-diffusion system* (2.1.1)–(2.1.4) *in the sense of the Definition 2.1.*

Finally in Section 2.4 we conclude with numerical simulations focusing on the behavior when the rescaled Debye length λ tends to 0.

2.2 A priori estimates

In order to prove the convergence of the scheme (2.1.13)–(2.1.19), we first need to prove a priori estimates on the approximate solution: $L^2(0,T,H^1)$ estimate on Ψ, $L^2(0,T,H^1)$ estimates on N and P and weak-BV inequalities on N and P. These estimates will provide compactness of the sequence of approximate solutions and permit to pass to the limit in the scheme. Such estimates have already been established for the implicit decoupled scheme proposed in [52, 54] by combining classical techniques developed for finite volume schemes for elliptic and parabolic equations (see [85]). However, as we are interested in this chapter in the quasi-neutral limit, we need a priori estimates which are independent of the rescaled Debye length λ and therefore, the techniques used in [52, 54] cannot directly apply. Like in the continuous context, the a priori estimates will be obtained as a consequence of an entropy estimate, with the control of the entropy production, similar to (2.1.12).

2.2.1 Discrete entropy estimate

Let us first introduce the discrete counterpart of the entropy functional and of the entropy dissipation introduced in [121] for instance and recalled at the end of Subsection 2.1.2.
The discrete entropy functional is defined by:

$$\begin{aligned}\mathbb{E}^n = &\sum_{K\in\mathcal{T}} \mathrm{m}(K)\left(H(N_K^n) - H(N_K^D) - \log(N_K^D)\left(N_K^n - N_K^D\right)\right)\\ &+ \sum_{K\in\mathcal{T}} \mathrm{m}(K)\left(H(P_K^n) - H(P_K^D) - \log(P_K^D)\left(P_K^n - P_K^D\right)\right)\\ &+ \frac{\lambda^2}{2}\left|\Psi^n - \Psi^D\right|_{1,\Omega}^2,\ \forall n\in\mathbb{N},\end{aligned}$$

where the discrete H^1-seminorm is defined by

$$|X|_{1,\Omega}^2 = \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\ (D_\sigma X)^2,\quad \forall X = ((X_K)_{K\in\mathcal{T}}, (X_\sigma^D)_{\sigma\in\mathcal{E}_{ext}^D}).$$

The discrete version of the entropy dissipation is

$$\begin{aligned}\mathbb{I}^n = &\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma \min\left(N_K^n, N_{K,\sigma}^n\right)\left(D_\sigma\left(\log N - \Psi\right)^n\right)^2\\ &+ \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma \min\left(P_K^n, P_{K,\sigma}^n\right)\left(D_\sigma\left(\log P + \Psi\right)^n\right)^2,\ \forall n\in\mathbb{N}.\end{aligned}$$

Proposition 2.2.1 (Control of entropy and entropy dissipation)**.** *Let us assume* (2.1.5), (2.1.6) *and that the solution to the numerical scheme* (2.1.13)–(2.1.19) *verifies the* L^∞*-estimates* (2.1.20)*. Then, there exists a constant* C_E *depending only on the data* Ω, T, m, M, N^D, P^D, Ψ^D, N^0, P^0, *such that for all* $\lambda > 0$,

$$\frac{\mathbb{E}^{n+1} - \mathbb{E}^n}{\Delta t} + \frac{1}{2}\,\mathbb{I}^{n+1} \le C_E. \tag{2.2.1}$$

Furthermore, if N^0 and P^0 satisfy the quasi-neutrality assumption (2.1.22)*, we have*

$$\sum_{n=0}^{N_T-1} \Delta t\, \mathbb{I}^{n+1} \leq C_E(1+\lambda^2). \tag{2.2.2}$$

Remark 1. Let us note that the estimate (2.2.2) depends on λ. However, this is not a problem to study the quasineutral limit $\lambda \to 0$ (we can assume for example that $\lambda \leq 1$, and then obtain an estimate independent of λ).

Proof. We adapt here the proof of Proposition 1.4.1 done in Chapter 1 for the study of the long-time behavior of the scheme (in this case, the entropy functional is defined relatively to the thermal equilibrium).
As H is a convex function, we have $\mathbb{E}^n \geq 0$ and $\mathbb{E}^{n+1} - \mathbb{E}^n \leq T_1 + T_2 + T_3$, with

$$\begin{aligned}
T_1 &= \sum_{K\in\mathcal{T}} \mathrm{m}(K) \left(\log\left(N_K^{n+1}\right) - \log\left(N_K^D\right)\right)\left(N_K^{n+1} - N_K^n\right),\\
T_2 &= \sum_{K\in\mathcal{T}} \mathrm{m}(K) \left(\log\left(P_K^{n+1}\right) - \log\left(P_K^D\right)\right)\left(P_K^{n+1} - P_K^n\right),\\
T_3 &= \frac{\lambda^2}{2}\left|\Psi^{n+1} - \Psi^D\right|_{1,\Omega}^2 - \frac{\lambda^2}{2}\left|\Psi^n - \Psi^D\right|_{1,\Omega}^2.
\end{aligned}$$

But, multiplying the scheme on N by $\Delta t\left(\log\left(N_K^{n+1}\right) - \log\left(N_K^D\right)\right)$, summing over $K \in \mathcal{T}$ and remarking that $\log(N_\sigma^{n+1}) = \log(N_\sigma^D)$ for all $\sigma \in \mathcal{E}_{ext}^D$, we get :

$$\begin{aligned}
T_1 &= -\Delta t \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} \mathcal{F}_{K,\sigma}^{n+1}\left(\log\left(N_K^{n+1}\right) - \log\left(N_K^D\right)\right)\\
&= \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \mathcal{F}_{K,\sigma}^{n+1}\left((D\log N)_{K,\sigma}^{n+1} - (D\log N)_{K,\sigma}^D\right).
\end{aligned} \tag{2.2.3}$$

Multiplying similarly the scheme on P by $\Delta t\left(\log\left(P_K^{n+1}\right) - \log\left(P_K^D\right)\right)$ and summing over $K \in \mathcal{T}$, we also get :

$$T_2 = \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \mathcal{G}_{K,\sigma}^{n+1}\left((D\log P)_{K,\sigma}^{n+1} - (D\log P)_{K,\sigma}^D\right). \tag{2.2.4}$$

Now, in order to estimate T_3, we make the difference between the scheme on Ψ written at two consecutive time steps. It implies

$$-\lambda^2 \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma D\Psi_{K,\sigma}^{n+1} + \lambda^2 \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma D\Psi_{K,\sigma}^n = \mathrm{m}(K)\Big((P_K^{n+1} - P_K^n) - (N_K^{n+1} - N_K^n)\Big).$$

Thanks to the schemes on N and P, it can be rewritten

$$-\lambda^2 \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\, D(\Psi^{n+1} - \Psi^D)_{K,\sigma} + \lambda^2 \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\, D(\Psi^n - \Psi^D)_{K,\sigma} = -\Delta t \sum_{\sigma\in\mathcal{E}_K} \left(\mathcal{G}_{K,\sigma}^{n+1} - \mathcal{F}_{K,\sigma}^{n+1}\right).$$

Multiplying this equality by $\Psi_K^{n+1} - \Psi_K^D$, summing over $K \in \mathcal{T}$, integrating by parts and using the boundary conditions, we obtain:

$$\lambda^2 \sum_{\sigma \in \mathcal{E}_{int} \cup \mathcal{E}_{ext}^D} \tau_\sigma \left(D_\sigma(\Psi^{n+1} - \Psi^D) \right)^2 - \lambda^2 \sum_{\sigma \in \mathcal{E}_{int} \cup \mathcal{E}_{ext}^D} \tau_\sigma \, D(\Psi^n - \Psi^D)_{K,\sigma} \, D(\Psi^{n+1} - \Psi^D)_{K,\sigma}$$
$$= \Delta t \sum_{\sigma \in \mathcal{E}_{int} \cup \mathcal{E}_{ext}^D} \left(\mathcal{G}_{K,\sigma}^{n+1} - \mathcal{F}_{K,\sigma}^{n+1} \right) D \left(\Psi^{n+1} - \Psi^D \right)_{K,\sigma}.$$

But, for all $K \in \mathcal{T}$ and all $\sigma \in \mathcal{E}_{K,int} \cup \mathcal{E}_{K,ext}^D$, we have

$$-D(\Psi^n - \Psi^D)_{K,\sigma} \, D(\Psi^{n+1} - \Psi^D)_{K,\sigma} \geq -\frac{1}{2} \left(D_\sigma(\Psi^n - \Psi^D) \right)^2 - \frac{1}{2} \left(D_\sigma(\Psi^{n+1} - \Psi^D) \right)^2,$$

and therefore for all $\lambda > 0$

$$T_3 \leq \Delta t \sum_{\sigma \in \mathcal{E}_{int} \cup \mathcal{E}_{ext}^D} (\mathcal{G}_{K,\sigma}^{n+1} - \mathcal{F}_{K,\sigma}^{n+1})(D\Psi_{K,\sigma}^{n+1} - D\Psi_{K,\sigma}^D). \tag{2.2.5}$$

From (2.2.3), (2.2.4) and (2.2.5), we get

$$\begin{aligned} \mathbb{E}^{n+1} - \mathbb{E}^n \leq \; & \Delta t \sum_{\sigma \in \mathcal{E}_{int} \cup \mathcal{E}_{ext}^D} \mathcal{F}_{K,\sigma}^{n+1} \left(D(\log N - \Psi)_{K,\sigma}^{n+1} - D(\log N - \Psi)_{K,\sigma}^D \right) \\ & + \Delta t \sum_{\sigma \in \mathcal{E}_{int} \cup \mathcal{E}_{ext}^D} \mathcal{G}_{K,\sigma}^{n+1} \left(D(\log P + \Psi)_{K,\sigma}^{n+1} - D(\log P + \Psi)_{K,\sigma}^D \right). \end{aligned}$$

Thanks to the inequalities (1.3.5) and (1.3.6) proved in Section 1.3 of Chapter 1, we have

$$\begin{aligned} \sum_{\sigma \in \mathcal{E}_{int} \cup \mathcal{E}_{ext}^D} \mathcal{F}_{K,\sigma}^{n+1} D(\log N - \Psi)_{K,\sigma}^{n+1} + \sum_{\sigma \in \mathcal{E}_{int} \cup \mathcal{E}_{ext}^D} \mathcal{G}_{K,\sigma}^{n+1} D(\log P + \Psi)_{K,\sigma}^{n+1} \leq \\ - \sum_{\sigma \in \mathcal{E}_{int} \cup \mathcal{E}_{ext}^D} \tau_\sigma \min(N_K^{n+1}, N_{K,\sigma}^{n+1}) \left(D_\sigma(\log N - \Psi)^{n+1} \right)^2 \\ - \sum_{\sigma \in \mathcal{E}_{int} \cup \mathcal{E}_{ext}^D} \tau_\sigma \min(P_K^{n+1}, P_{K,\sigma}^{n+1}) \left(D_\sigma(\log P + \Psi)^{n+1} \right)^2. \end{aligned}$$

Using now (1.3.7), (1.3.8), Cauchy-Schwarz and Young's inequalities, we get

$$\begin{aligned}\left|\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D}\mathcal{F}_{K,\sigma}^{n+1}D(\log N-\Psi)_{K,\sigma}^D\right|+\left|\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D}\mathcal{G}_{K,\sigma}^{n+1}D(\log P+\Psi)_{K,\sigma}^D\right|\leq\\ \frac{1}{2}\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D}\tau_\sigma\min(N_K^{n+1},N_{K,\sigma}^{n+1})\left(D_\sigma(\log N-\Psi)^{n+1}\right)^2\\ +\frac{1}{2}\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D}\tau_\sigma\min(P_K^{n+1},P_{K,\sigma}^{n+1})\left(D_\sigma(\log P+\Psi)^{n+1}\right)^2\\ +\frac{1}{2}\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D}\tau_\sigma\frac{\max(N_K^{n+1},N_{K,\sigma}^{n+1})^2}{\min(N_K^{n+1},N_{K,\sigma}^{n+1})}\left(D_\sigma(\log N-\Psi)^D\right)^2\\ +\frac{1}{2}\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D}\tau_\sigma\frac{\max(P_K^{n+1},P_{K,\sigma}^{n+1})^2}{\min(P_K^{n+1},P_{K,\sigma}^{n+1})}\left(D_\sigma(\log P+\Psi)^D\right)^2.\end{aligned}$$

Finally, using the L^∞-estimates (2.1.20),we obtain

$$\begin{aligned}\frac{\mathbb{E}^{n+1}-\mathbb{E}^n}{\Delta t}\leq&-\frac{1}{2}\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D}\tau_\sigma\min(N_K^{n+1},N_L^{n+1})\left(D_\sigma(\log N-\Psi)^{n+1}\right)^2\\ &-\frac{1}{2}\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D}\tau_\sigma\min(P_K^{n+1},P_L^{n+1})\left(D_\sigma(\log P+\Psi)^{n+1}\right)^2\\ &+\frac{M^2}{2m}\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D}\tau_\sigma\left(D_\sigma(\log N-\Psi)^D\right)^2\\ &+\frac{M^2}{2m}\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D}\tau_\sigma\left(D_\sigma(\log P+\Psi)^D\right)^2,\end{aligned}$$

which rewrites

$$\begin{aligned}&\frac{\mathbb{E}^{n+1}-\mathbb{E}^n}{\Delta t}+\frac{1}{2}\,\mathbb{I}^{n+1}\leq\\ &\qquad\frac{M^2}{2m}\left(\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D}\tau_\sigma\left(D_\sigma(\log N-\Psi)^D\right)^2+\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D}\tau_\sigma\left(D_\sigma(\log P+\Psi)^D\right)^2\right).\end{aligned}$$

But, as the boundary data N^D, P^D and Ψ^D belong to $H^1(\Omega)$ (hypothesis (2.1.6)), the right hand side is bounded by a constant $\mathcal{K}$ depending only on Ω, m, M, N^D, P^D and Ψ^D, which yields (2.2.1).
Summing over $n\in\{0,\dots N_T-1\}$, we get

$$\sum_{n=0}^{N_T-1}\Delta t\,\mathbb{I}^{n+1}\ \leq\ \mathbb{E}^{N_T}+\sum_{n=0}^{N_T-1}\Delta t\,\mathbb{I}^{n+1}\ \leq\ T\,\mathcal{K}+\mathbb{E}^0.$$

It remains now to bound $\mathbb{E}^0$. As the function H satisfies the following inequality:

$$\forall x,\, y > 0, \quad H(y) - H(x) - \log(x)(y-x) \leq \frac{1}{\min(x,y)}\,\frac{(y-x)^2}{2},$$

we get using (2.1.8):

$$\sum_{K\in\mathcal{T}} \mathrm{m}(K)\left(H(N_K^0) - H(N_K^D) - \log(N_K^D)(N_K^0 - N_K^D)\right) \leq \mathrm{m}(\Omega)\,\frac{(M-m)^2}{2\,m},$$

$$\sum_{K\in\mathcal{T}} \mathrm{m}(K)\left(H(P_K^0) - H(P_K^D) - \log(P_K^D)(P_K^0 - P_K^D)\right) \leq \mathrm{m}(\Omega)\,\frac{(M-m)^2}{2\,m}.$$

Then multiplying the scheme on Ψ at $n=0$ by $\Psi_K^0 - \Psi_K^D$ and summing over $K\in\mathcal{T}$, we get

$$\lambda^2 \sum_{\sigma\in\mathcal{E}} \tau_\sigma\, D\Psi_{K,\sigma}^0 \left(D\Psi_{K,\sigma}^0 - D\Psi_{K,\sigma}^D\right) = \sum_{K\in\mathcal{T}} \mathrm{m}(K)(P_K^0 - N_K^0)(\Psi_K^0 - \Psi_K^D) = 0$$

if the initial conditions satisfy the quasi-neutrality assumption (2.1.22).
Then, as $a(a-b) \geq (a-b)^2/2 - b^2/2$ for all $a,\, b\in\mathbb{R}$, we obtain

$$\frac{\lambda^2}{2}\left|\Psi^0 - \Psi^D\right|_{1,\Omega}^2 \leq \frac{\lambda^2}{2}\left|\Psi^D\right|_{1,\Omega}^2.$$

Finally using (2.1.6) we have

$$\mathbb{E}^0 \leq \mathcal{K}\,(1+\lambda^2),$$

where $\mathcal{K}$ depends only on N^0, P^0, N^D, P^D, Ψ^D and not on $\lambda > 0$, which gives (2.2.2). □

Remark 2. Let us just note that we do not assume in Proposition 2.2.1 that the doping profile vanishes. Indeed, the terms due to the doping profile vanishes when we estimate T_3. However, in order to define the discrete entropy and the discrete dissipation and then to prove (2.2.1) and (2.2.2), we need that the approximate densities are strictly positive and satisfy (2.1.20). When the doping profile vanishes, it is guaranteed by Theorem 2.1.1.

2.2.2 Weak BV-inequalities on N and P

We will now establish weak-BV inequalities on N and P.

Proposition 2.2.2. *Let us assume* (2.1.5), (2.1.6), (2.1.8) *and* $C = 0$. *We further assume that the boundary and initial conditions satisfy the quasi-neutrality assumption* (2.1.21)–(2.1.22). *Then, there exists a constant* C_{BV} *depending only on* Ω, T, m, M, N^D, P^D, Ψ^D, N^0, P^0, *such that, for all* $\lambda > 0$, *a solution to the scheme defined by* (2.1.13)–(2.1.19) *verifies*

$$\sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma \left(D_\sigma P^{n+1}\right)^2 D_\sigma\Psi^{n+1} + \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma \left(D_\sigma N^{n+1}\right)^2 D_\sigma\Psi^{n+1} \leq C_{BV}. \tag{2.2.6}$$

Proof. Let us set

$$T_{BV} = \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma \left(D_\sigma P^{n+1}\right)^2 D_\sigma \Psi^{n+1} + \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma \left(D_\sigma N^{n+1}\right)^2 D_\sigma \Psi^{n+1}.$$

We follow the ideas of [54]. We multiply the scheme on N by $\Delta t(N_K^{n+1} - N_K^D)$ and the scheme on P by $\Delta t(P_K^{n+1} - P_K^D)$ and we sum over $K \in \mathcal{T}$ and n. It yields

$$E_1 + E_2 + E_3 + F_1 + F_2 + F_3 = 0, \tag{2.2.7}$$

with

$$E_1 = \sum_{n=0}^{N_T-1} \sum_{K\in\mathcal{T}} \mathrm{m}(K)(N_K^{n+1} - N_K^n)(N_K^{n+1} - N_K^D),$$
$$F_1 = \sum_{n=0}^{N_T-1} \sum_{K\in\mathcal{T}} \mathrm{m}(K)(P_K^{n+1} - P_K^n)(P_K^{n+1} - P_K^D),$$
$$E_2 = -\sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \mathcal{F}_{K,\sigma}^{n+1} D N_{K,\sigma}^{n+1}, \quad F_2 = -\sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \mathcal{G}_{K,\sigma}^{n+1} D P_{K,\sigma}^{n+1},$$
$$E_3 = \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \mathcal{F}_{K,\sigma}^{n+1} D N_{K,\sigma}^{D}, \quad F_3 = \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \mathcal{G}_{K,\sigma}^{n+1} D P_{K,\sigma}^{D}.$$

As it is classical, we have

$$E_1 \geq -\frac{1}{2}\sum_{K\in\mathcal{T}} \mathrm{m}(K)(N_K^0 - N_K^D)^2 \text{ and } F_1 \geq -\frac{1}{2}\sum_{K\in\mathcal{T}} \mathrm{m}(K)(P_K^0 - P_K^D)^2.$$

Hence,

$$E_1 \geq -\frac{1}{2}\,\mathrm{m}(\Omega)\,(M-m)^2 \text{ and } F_1 \geq -\frac{1}{2}\,\mathrm{m}(\Omega)\,(M-m)^2. \tag{2.2.8}$$

We may also bound the terms E_3 and F_3. Indeed, using successively the property of the

flux $\mathcal{F}_{K,\sigma}^{n+1}$ (1.3.7), the L^∞ estimates (2.1.20) and Cauchy-Schwarz inequality, we get

$$
\begin{aligned}
|E_3| &\leq \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma \max(N_K^{n+1}, N_{K,\sigma}^{n+1})\, D_\sigma(\log N - \Psi)^{n+1}\, D_\sigma N^D \\
&\leq \frac{M}{\sqrt{m}} \left(\sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma \min(N_K^{n+1}, N_{K,\sigma}^{n+1}) \left(D_\sigma(\log N - \Psi)^{n+1} \right)^2 \right)^{1/2} \\
&\quad \times \left(\sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma \left(D_\sigma N^D \right)^2 \right)^{1/2}.
\end{aligned}
$$

But the right-hand side is bounded thanks to the control of the entropy dissipation (2.2.2) and the hypothesis (2.1.6). Following similar computations for F_3, we get

$$
|E_3| \leq \mathcal{K} \text{ and } |F_3| \leq \mathcal{K} \tag{2.2.9}
$$

with $\mathcal{K}$ depending only on C_E, M, m, N^D and P^D and independent of $\lambda > 0$.
We focus now on the main terms E_2 and F_2. Using the definition (2.1.19) of the Bernoulli function, the numerical fluxes $\mathcal{F}_{K,\sigma}^{n+1}$ and $\mathcal{G}_{K,\sigma}^{n+1}$, defined by (2.1.17) and (2.1.18), rewrite:

$$
\begin{aligned}
\mathcal{F}_{K,\sigma}^{n+1} &= \frac{\tau_\sigma}{2} \left[D\Psi_{K,\sigma}^{n+1}(N_K^{n+1} + N_{K,\sigma}^{n+1}) + D\Psi_{K,\sigma}^{n+1} \coth\left(\frac{D\Psi_{K,\sigma}^{n+1}}{2} \right) (N_K^{n+1} - N_{K,\sigma}^{n+1}) \right], \\
\mathcal{G}_{K,\sigma}^{n+1} &= \frac{\tau_\sigma}{2} \left[-D\Psi_{K,\sigma}^{n+1}(P_K^{n+1} + P_{K,\sigma}^{n+1}) + D\Psi_{K,\sigma}^{n+1} \coth\left(\frac{D\Psi_{K,\sigma}^{n+1}}{2} \right) (P_K^{n+1} - P_{K,\sigma}^{n+1}) \right].
\end{aligned}
$$

As $x\coth(x) \geq |x|$ for all $x \in \mathbb{R}$, we obtain

$$
\begin{aligned}
E_2 \geq & -\frac{1}{2} \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\, D\Psi_{K,\sigma}^{n+1}((N_{K,\sigma}^{n+1})^2 - (N_K^{n+1})^2) \\
& + \frac{1}{2} \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\, D_\sigma\Psi^{n+1} \left(D_\sigma N^{n+1} \right)^2, \\
F_2 \geq & \frac{1}{2} \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\, D\Psi_{K,\sigma}^{n+1}((P_{K,\sigma}^{n+1})^2 - (P_K^{n+1})^2) \\
& + \frac{1}{2} \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\, D_\sigma\Psi^{n+1} \left(D_\sigma P^{n+1} \right)^2.
\end{aligned}
$$

Summing these two inequalities, we can then integrate by parts due to the quasi-neutrality

of the boundary conditions (2.1.21) and get

$$E_2+F_2 \geq \frac{1}{2}\sum_{n=0}^{N_T-1}\Delta t\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}\tau_\sigma\, D\Psi_{K,\sigma}^{n+1}\Big((N_K^{n+1})^2-(P_K^{n+1})^2\Big)+\frac{1}{2}T_{BV}$$
$$\geq \frac{1}{2\lambda^2}\sum_{n=0}^{N_T-1}\Delta t\sum_{K\in\mathcal{T}}\mathrm{m}(K)(N_K^{n+1}-P_K^{n+1})\Big((N_K^{n+1})^2-(P_K^{n+1})^2\Big)+\frac{1}{2}T_{BV},$$

thanks to the scheme on Ψ. As the function $x\mapsto x^2$ is nondecreasing on $\mathbb{R}^+$, it yields

$$E_2+F_2 \geq \frac{1}{2}T_{BV}. \tag{2.2.10}$$

Finally, we deduce the weak-BV inequality (2.2.6) from (2.2.7), (2.2.8), (2.2.9) and (2.2.10). □

2.2.3 Discrete $L^2(0,T,H^1(\Omega))$ estimate on Ψ

Let us now turn on the discrete $L^2(0,T,H^1(\Omega))$ estimate on Ψ. Once more, we will use Proposition 2.2.1 in the proof.

Proposition 2.2.3. *Let us assume* (2.1.5), (2.1.6), (2.1.8), $C=0$ *and the quasi-neutrality assumptions* (2.1.21)–(2.1.22). *Then, there exists* C_Ψ *depending only on* Ω, T, m, M, N^D, P^D, Ψ^D, N^0, P^0, *such that, for all* $\lambda>0$,

$$\sum_{n=0}^{N_T-1}\Delta t\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D}\tau_\sigma\left(D_\sigma\Psi^{n+1}\right)^2\leq C_\Psi. \tag{2.2.11}$$

Proof. We follow here ideas developed by I. Gasser in [94] at the continuous level in the case of homogeneous boundary conditions. Let us set

$$\mathcal{J}=\sum_{n=0}^{N_T-1}\Delta t\sum_{K\in\mathcal{T}}\mathrm{m}(K)\frac{(N_K^{n+1}-P_K^{n+1})^2}{\lambda^2} \tag{2.2.12}$$
$$+\sum_{n=0}^{N_T-1}\Delta t\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D}\tau_\sigma\left(\min(N_K^{n+1},N_{K,\sigma}^{n+1})+\min(P_K^{n+1},P_{K,\sigma}^{n+1})\right)\left(D_\sigma\Psi^{n+1}\right)^2.$$

Multipliying the scheme on Ψ by $\Delta t\,(P_K^{n+1}-N_K^{n+1})/\lambda^2$ and summing over $K\in\mathcal{T}$ and $n\in\{0,\dots,N_T-1\}$, we get

$$\sum_{n=0}^{N_T-1}\sum_{K\in\mathcal{T}}\mathrm{m}(K)\frac{(N_K^{n+1}-P_K^{n+1})^2}{\lambda^2}=-\sum_{n=0}^{N_T-1}\Delta t\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}\tau_\sigma\, D\Psi_{K,\sigma}^{n+1}(P_K^{n+1}-N_K^{n+1})$$
$$=\sum_{n=0}^{N_T-1}\Delta t\sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D}\tau_\sigma\, D\Psi_{K,\sigma}^{n+1}(DP_{K,\sigma}^{n+1}-DN_{K,\sigma}^{n+1}),$$

due to the quasi-neutrality of the boundary conditions (2.1.21). Therefore, $\mathcal{J}$ rewrites :

$$\mathcal{J} = \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\, D\Psi_{K,\sigma}^{n+1} \left(\min(N_K^{n+1}, N_{K,\sigma}^{n+1})\, D\Psi_{K,\sigma}^{n+1} - DN_{K,\sigma}^{n+1}\right)$$
$$+ \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\, D\Psi_{K,\sigma}^{n+1} \left(\min(P_K^{n+1}, P_{K,\sigma}^{n+1})\, D\Psi_{K,\sigma}^{n+1} + DP_{K,\sigma}^{n+1}\right).$$

It may be splitted into $\mathcal{J} = \mathcal{J}_1 + \mathcal{J}_2$ with

$$\mathcal{J}_1 = \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\, D\Psi_{K,\sigma}^{n+1} \min(P_K^{n+1}, P_{K,\sigma}^{n+1})\, D(\log P + \Psi)_{K,\sigma}^{n+1}$$
$$- \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\, D\Psi_{K,\sigma}^{n+1} \min(N_K^{n+1}, N_{K,\sigma}^{n+1})\, D(\log N - \Psi)_{K,\sigma}^{n+1},$$
$$\mathcal{J}_2 = \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\, D\Psi_{K,\sigma}^{n+1} \left(DP_{K,\sigma}^{n+1} - \min(P_K^{n+1}, P_{K,\sigma}^{n+1})\, D(\log P)_{K,\sigma}^{n+1}\right)$$
$$- \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\, D\Psi_{K,\sigma}^{n+1} \left(DN_{K,\sigma}^{n+1} - \min(N_K^{n+1}, P_{K,\sigma}^{n+1})\, D(\log N)_{K,\sigma}^{n+1}\right).$$

Applying successively Cauchy-Schwarz and Young inequality on $\mathcal{J}_1$, we get

(2.2.13)
$$|\mathcal{J}_1| \le \frac{1}{2} \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma \left(D_\sigma \Psi^{n+1}\right)^2 \left(\min(N_K^{n+1}, N_{K,\sigma}^{n+1}) + \min(P_K^{n+1}, P_{K,\sigma}^{n+1})\right)$$
$$+ \frac{1}{2} \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma \min(P_K^{n+1}, P_{K,\sigma}^{n+1}) \left(D_\sigma(\log P + \Psi)^{n+1}\right)^2$$
$$+ \frac{1}{2} \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma \min(N_K^{n+1}, N_{K,\sigma}^{n+1}) \left(D_\sigma(\log N - \Psi)^{n+1}\right)^2$$
$$\le \frac{1}{2} \sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma \left(D_\sigma \Psi^{n+1}\right)^2 \left(\min(N_K^{n+1}, N_{K,\sigma}^{n+1}) + \min(P_K^{n+1}, P_{K,\sigma}^{n+1})\right) + \frac{1}{2} C_E.$$

Let us now estimate the term $\mathcal{J}_2$ which does not appear at the continuous level because $\nabla N = N \nabla \log N$. For all $x, y > 0$ we have

$$\left|\log y - \log x - \frac{y-x}{\min(x,y)}\right| \le \frac{(x-y)^2}{2\min(x,y)^2}.$$

It yields

$$\left|DP_{K,\sigma}^{n+1} - \min(P_K^{n+1}, P_{K,\sigma}^{n+1}) D(\log P)_{K,\sigma}^{n+1}\right| \leq \frac{\left(D_\sigma P^{n+1}\right)^2}{2m},$$
$$\left|DN_{K,\sigma}^{n+1} - \min(N_K^{n+1}, N_{K,\sigma}^{n+1}) D(\log N)_{K,\sigma}^{n+1}\right| \leq \frac{\left(D_\sigma N^{n+1}\right)^2}{2m},$$

and

(2.2.14)
$$|\mathcal{J}_2| \leq \frac{1}{2\,m}\left(\sum_{n=0}^{N_T-1}\Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\, D_\sigma \Psi^{n+1}\,\left[\left(D_\sigma P^{n+1}\right)^2 + \left(D_\sigma N^{n+1}\right)^2\right]\right) \leq \frac{1}{2\,m}\, C_{BV},$$

thanks to the weak-BV inequality (2.2.6). From (2.2.12), (2.2.13) and (2.2.14), we get :

$$\begin{aligned}
&\sum_{n=0}^{N_T-1}\sum_{K\in\mathcal{T}} \mathrm{m}(K)\frac{(N_K^{n+1}-P_K^{n+1})^2}{\lambda^2} \\
&+\frac{1}{2}\sum_{n=0}^{N_T-1}\Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\,\left(\min(N_K^{n+1}, N_{K,\sigma}^{n+1}) + \min(P_K^{n+1}, P_{K,\sigma}^{n+1})\right)\left(D_\sigma\Psi^{n+1}\right)^2 \\
&\qquad \leq \frac{m\,C_E + C_{BV}}{2\,m}.
\end{aligned}$$

As N and P are lower bounded by m (2.1.20), it yields (2.2.11) and concludes the proof of Proposition 2.2.3. □

2.2.4 Discrete $L^2(0,T,H^1(\Omega))$ estimates on the densities

Proposition 2.2.4. *Let us assume* (2.1.5), (2.1.6), (2.1.8), $C = 0$ *and the quasi-neutrality assumptions* (2.1.21)–(2.1.22). *Then, there exists* C_D *depending only on* Ω, T, m, M, N^D, P^D, Ψ^D, N^0, P^0, *such that, for all* $\lambda > 0$,

$$\sum_{n=0}^{N_T-1}\Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\left(D_\sigma N^{n+1}\right)^2 + \sum_{n=0}^{N_T-1}\Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\left(D_\sigma P^{n+1}\right)^2 \leq C_D. \tag{2.2.15}$$

Proof. We start as in the proof of Proposition 2.2.2 with (2.2.7). But we treat in a different manner the terms E_2 and F_2. Indeed, for all $K \in \mathcal{T}$ and all $\sigma \in \mathcal{E}_{K,int} \cup \mathcal{E}_{K,ext}^D$, the Scharfetter-Gummel fluxes $\mathcal{F}_{K,\sigma}^{n+1}$ and $\mathcal{G}_{K,\sigma}^{n+1}$ defined by (2.1.17) and (2.1.18) rewrite

$$\mathcal{F}_{K,\sigma}^{n+1} = \tau_\sigma\left(-DN_{K,\sigma}^{n+1} + \tilde{B}(-D\Psi_{K,\sigma}^{n+1})N_K^{n+1} - \tilde{B}(D\Psi_{K,\sigma}^{n+1})N_{K,\sigma}^{n+1}\right) = -\tau_\sigma\, DN_{K,\sigma}^{n+1} + \widetilde{\mathcal{F}}_{K,\sigma}^{n+1},$$
$$\mathcal{G}_{K,\sigma}^{n+1} = \tau_\sigma\left(-DP_{K,\sigma}^{n+1} + \tilde{B}(D\Psi_{K,\sigma}^{n+1})P_K^{n+1} - \tilde{B}(-D\Psi_{K,\sigma}^{n+1})P_{K,\sigma}^{n+1}\right) = -\tau_\sigma\, DP_{K,\sigma}^{n+1} + \widetilde{\mathcal{G}}_{K,\sigma}^{n+1},$$

with $\tilde{B}$ defined by $\tilde{B}(x) = B(x) - 1$ for all $x \in \mathbb{R}$. Therefore

(2.2.16)
$$E_2 + F_2 = \sum_{n=0}^{N_T-1}\Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\left(D_\sigma N^{n+1}\right)^2 + \sum_{n=0}^{N_T-1}\Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \tau_\sigma\left(D_\sigma P^{n+1}\right)^2 + \tilde{E}_2 + \tilde{F}_2,$$

with

$$\tilde{E}_2 = -\sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \widetilde{\mathcal{F}}_{K,\sigma}^{n+1}\, DN_{K,\sigma}^{n+1} \text{ and } \tilde{F}_2 = -\sum_{n=0}^{N_T-1} \Delta t \sum_{\sigma\in\mathcal{E}_{int}\cup\mathcal{E}_{ext}^D} \widetilde{\mathcal{G}}_{K,\sigma}^{n+1}\, DP_{K,\sigma}^{n+1}.$$

As for the fluxes $\mathcal{F}_{K,\sigma}^{n+1}$, we can rewrite the fluxes $\widetilde{\mathcal{F}}_{K,\sigma}^{n+1}$ either under the form (1.3.3) or (1.3.4) with $\tilde{B}$ instead of B. Then, as $x(x-y) = \frac{1}{2}(x-y)^2 + \frac{1}{2}(x^2-y^2)$, we get either

$$\begin{aligned}(2.2.17)\quad -\widetilde{\mathcal{F}}_{K,\sigma}^{n+1}\, DN_{K,\sigma}^{n+1} = \tau_\sigma \Bigg(&\frac{D\Psi_{K,\sigma}^{n+1}}{2}(N_K^{n+1} - N_{K,\sigma}^{n+1})^2 \\ &+\frac{D\Psi_{K,\sigma}^{n+1}}{2}\left((N_K^{n+1})^2 - (N_{K,\sigma}^{n+1})^2\right) + \tilde{B}(D\Psi_{K,\sigma}^{n+1})\left(D_\sigma N^{n+1}\right)^2\Bigg)\end{aligned}$$

or

$$\begin{aligned}(2.2.18)\quad -\widetilde{\mathcal{F}}_{K,\sigma}^{n+1}\, DN_{K,\sigma}^{n+1} = \tau_\sigma \Bigg(&-\frac{D\Psi_{K,\sigma}^{n+1}}{2}(N_K^{n+1} - N_{K,\sigma}^{n+1})^2 \\ &-\frac{D\Psi_{K,\sigma}^{n+1}}{2}\left((N_{K,\sigma}^{n+1})^2 - (N_K^{n+1})^2\right) + \tilde{B}(-D\Psi_{K,\sigma}^{n+1})\left(D_\sigma N^{n+1}\right)^2\Bigg).\end{aligned}$$

But $\tilde{B}(x) \geq 0$ for all $x \leq 0$ and $\tilde{B}(-x) \geq 0$ for all $x \geq 0$. Then, using (2.2.17) when $D\Psi_{K,\sigma}^{n+1} \leq 0$ and (2.2.18) when $D\Psi_{K,\sigma}^{n+1} \geq 0$, we obtain that in both cases

$$-\widetilde{\mathcal{F}}_{K,\sigma}^{n+1}\, DN_{K,\sigma}^{n+1} \geq \frac{\tau_\sigma}{2}\left(-|D\Psi_{K,\sigma}^{n+1}|(DN_{K,\sigma}^{n+1})^2 + D\Psi_{K,\sigma}^{n+1}\left((N_K^{n+1})^2 - (N_{K,\sigma}^{n+1})^2\right)\right).$$

Similarly, we have

$$-\widetilde{\mathcal{G}}_{K,\sigma}^{n+1}\, DP_{K,\sigma}^{n+1} \geq \frac{\tau_\sigma}{2}\left(-|D\Psi_{K,\sigma}^{n+1}|(DP_{K,\sigma}^{n+1})^2 - D\Psi_{K,\sigma}^{n+1}\left((P_K^{n+1})^2 - (P_{K,\sigma}^{n+1})^2\right)\right).$$

It yields after a discrete integration by parts

$$\tilde{E}_2 + \tilde{F}_2 \geq -\frac{1}{2}\, T_{BV} + \frac{1}{2}\sum_{n=0}^{N_T-1} \Delta t \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma D\Psi_{K,\sigma}^{n+1}\left((N_K^{n+1})^2 - (P_K^{n+1})^2\right)$$

and

$$(2.2.19)\qquad \tilde{E}_2 + \tilde{F}_2 \geq -\frac{1}{2}\, T_{BV}$$

thanks to the scheme on Ψ. Then, we deduce the discrete $L^2(0,T,H^1)$ estimate (2.2.15) on N and P from (2.2.7), (2.2.8), (2.2.9), (2.2.16), (2.2.19) and the BV estimate (2.2.6). □

2.3 Convergence of the scheme

In this section, we prove Theorem 2.1.2, which gives the convergence of the approximate solution $(N_\delta, P_\delta, \Psi_\delta)$ toward a weak solution (N, P, Ψ) of the drift-diffusion system (2.1.1)–(2.1.4) for all $\lambda > 0$. We begin classically with compactness results, deduced from the a priori estimates proved in the previous section, and then we pass to the limit in the scheme. The compactness results can be detailed as in Subsection 3.4.1 of Chapter 3.

Using the discrete $L^2(0,T;H^1(\Omega))$ estimates on N and P, we can write a result similar to Lemma 3.4.1 of Chapter 3, which gives space and time translate estimates of N_δ and P_δ. Then the following result is proved in [52]:

Lemma 2.3.1 (Compactness)**.** *Let us assume* (2.1.5), (2.1.6), (2.1.8), $C = 0$ *and the quasi-neutrality assumptions* (2.1.21)–(2.1.22)*. Then there exists* N, $P \in L^2(0,T;H^1(\Omega))$ *and* $\Psi \in \left(L^2(0,T;H^1(\Omega))\right)^d$ *such that, up to subsequences,*

$$N_\delta \to N \text{ and } P_\delta \to P \text{ in } L^2(\Omega\times(0,T)) \text{ strongly, as } \delta \to 0,$$

$$\Psi_\delta \rightharpoonup \Psi \text{ in } \left(L^2(\Omega\times(0,T))\right)^d \text{ weakly, as } \delta \to 0,$$

$$dN_\delta \rightharpoonup \nabla N,\ dP_\delta \rightharpoonup \nabla P \text{ and } d\Psi_\delta \rightharpoonup \nabla\Psi \text{ in } \left(L^2(\Omega\times(0,T))\right)^d \text{ weakly, as } \delta \to 0.$$

Now it remains to prove that the functions N, P, Ψ defined in Lemma 2.3.1 satisfy Definition 2.1 of the weak solution. The result is given in Theorem 2.3.1. Passing to the limit in the Poisson equation can be done exactly as in [52, Theorem 5.1]. The main difficulty in proving Theorem 2.3.1 comes from the fact that the diffusive and convective terms are put together in the Scharfetter-Gummel flux. The detailed proof will be achieved in Theorem 3.4.1 of Chapter 3 for the generalization of the Scharfetter-Gummel fluxes in the case of a nonlinear diffusion. We only give here the main steps of the proof.

Theorem 2.3.1. *Assume* (2.1.5), (2.1.6), (2.1.8), $C = 0$ *and the quasi-neutrality assumptions* (2.1.21)–(2.1.22) *hold. Then* (N, P, Ψ) *defined in Lemma 2.3.1 is a weak solution to the problem* (2.1.1)–(2.1.4) *in the sense of Definition 2.1.*

Proof. As explained above, we apply the proof of Theorem 3.4.1 in Chapter 3 with $r(s) = s$ (linear diffusion) and with $q_{K,\sigma} = D\Psi_{K,\sigma}/d_\sigma$. The only difference is that a discrete $L^\infty(\Omega\times(0,T))$ estimate of $\mathbf{q}$ is used in Chapter 3, and we do not have such an estimate for $d\Psi_\delta$. We only focus on (2.1.9) since the proof of (2.1.10) is almost the same. Concerning the proof of the boundary conditions $N - N^D$, $P - P^D \in L^2(0,T;V)$, it is done in [52]. Let $\phi \in \mathcal{D}(\Omega\times[0,T))$ and $\delta > 0$ be small enough such that $\operatorname{supp}\phi \subset [0,(N_T-1)\Delta t)\times\{x \in \Omega : \mathrm{d}(x,\partial\Omega) > \delta\}$. We use the same notations as in the proof of Theorem 3.4.1 in

Chapter 3: we define

$$B_{10}(\delta) = -\int_0^T \int_\Omega N_\delta \, \partial_t \phi \, dx \, dt - \int_\Omega N_\delta(x,0) \, \phi(x,0) \, dx,$$
$$B_{20}(\delta) = \int_0^T \int_\Omega dN_\delta \cdot \nabla \phi \, dx \, dt,$$
$$B_{30}(\delta) = -\int_0^T \int_\Omega N_\delta \, d\Psi_\delta \cdot \nabla \phi \, dx \, dt,$$

and

$$\varepsilon(\delta) = -\left[B_{10}(\delta) + B_{20}(\delta) + B_{30}(\delta)\right].$$

Let $\phi_K^n := \phi(t^n, x_K)$ for all $K \in \mathcal{T}$ and $n = 0, ..., N_T$. Multiplying the scheme on N by $\Delta t \, \phi_K^n$ and summing for K and n, we obtain

$$B_1(\delta) + B_2(\delta) + B_3(\delta) = 0,$$

where

$$B_1(\delta) = \sum_{n=0}^{N_T-1} \sum_{K \in \mathcal{T}} \mathrm{m}(K) \left(N_K^{n+1} - N_K^n\right) \phi_K^n,$$
$$B_2(\delta) = \sum_{n=0}^{N_T-1} \Delta t \sum_{K \in \mathcal{T}} \sum_{\substack{\sigma \in \mathcal{E}_{int} \\ \sigma = K|L}} \tau_\sigma \frac{D\Psi_{K,\sigma}^{n+1}}{2} \coth\left(\frac{D\Psi_{K,\sigma}^{n+1}}{2}\right) \left(N_L^{n+1} - N_K^{n+1}\right) \phi_K^n,$$
$$B_3(\delta) = \sum_{n=0}^{N_T-1} \Delta t \sum_{K \in \mathcal{T}} \sum_{\substack{\sigma \in \mathcal{E}_{int} \\ \sigma = K|L}} \tau_\sigma \frac{D\Psi_{K,\sigma}^{n+1}}{2} \left(N_K^{n+1} + N_L^{n+1}\right) \phi_K^n.$$

From the strong convergence of $(N_\delta)_{\delta>0}$ in $L^2(\Omega \times (0,T))$ and the weak convergence of $(dN_\delta)_{\delta>0}$ and $(d\Psi_\delta)_{\delta>0}$ in $\left(L^2(\Omega \times (0,T))\right)^d$, it is easy to see that

$$\varepsilon(\delta) \to \int_0^T \int_\Omega (N \, \partial_t \phi - \nabla N \cdot \nabla \phi + N \, \nabla \Psi \cdot \nabla \phi) \, dx \, dt + \int_\Omega N_0(x) \phi(x,0) \, dx \quad \text{as} \quad \delta \to 0.$$

Therefore it suffices to prove that $\varepsilon(\delta) \to 0$ as $\delta \to 0$ and to this end we prove that $B_j(\delta) - B_{j0}(\delta) \to 0$ as $\delta \to 0$, $j = 1, 2, 3$. The term $B_1(\delta) - B_{10}(\delta)$ is treated in [52, Theorem 5.2]. Then we treat the term $B_2(\delta) - B_{20}(\delta)$ like in Theorem 3.4.1 of Chapter 3. We write $B_2(\delta) = B_{21}(\delta) + B_{22}(\delta)$, with

$$B_{21}(\delta) = \sum_{n=0}^{N_T-1} \Delta t \sum_{\substack{\sigma \in \mathcal{E}_{int} \\ \sigma = K|L}} \tau_\sigma \left(N_L^{n+1} - N_K^{n+1}\right) \left(\phi_L^n - \phi_K^n\right),$$
$$B_{22}(\delta) = \sum_{n=0}^{N_T-1} \Delta t \sum_{\substack{\sigma \in \mathcal{E}_{int} \\ \sigma = K|L}} \tau_\sigma \left(\frac{D\Psi_{K,\sigma}^{n+1}}{2} \coth\left(\frac{D\Psi_{K,\sigma}^{n+1}}{2}\right) - 1\right) \left(N_L^{n+1} - N_K^{n+1}\right) \left(\phi_L^n - \phi_K^n\right).$$

Using the definition of dN_δ, we rewrite $B_{20}(\delta)$ as

$$B_{20}(\delta) = B_{210}(\delta) = \sum_{n=0}^{N_T-1} \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \frac{\mathrm{m}(\sigma)}{\mathrm{m}(\mathrm{T_{K,\sigma}})} \left(N_L^{n+1} - N_K^{n+1}\right) \int_{t^n}^{t^{n+1}} \int_{T_{K,\sigma}} \nabla\phi(x,t)\cdot \mathbf{n}_{K,\sigma}\, dx\, dt.$$

Using the same computations as in Theorem 3.4.1, we deduce from the $L^2(0,T;H^1(\Omega))$ estimate (2.2.15) on N that

$$B_{21}(\delta) - B_{210}(\delta) \to 0 \text{ as } \delta \to 0.$$

Concerning $B_{22}(\delta)$, since $x \mapsto x\coth(x)$ is a 1-Lipschitz continuous function and is equal to 1 in 0, we have

$$|B_{22}(\delta)| \le \sum_{n=0}^{N_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \frac{\tau_\sigma}{2}\, D_\sigma\Psi^{n+1}\, D_\sigma N^{n+1}\, |\phi_L^n - \phi_K^n|\,.$$

Then using the regularity of ϕ, the Cauchy-Schwarz inequality and the $L^2(0,T;H^1)$ estimates (2.2.11) and (2.2.15) on Ψ and N respectively, there exists $C>0$ only depending on Ω and T such that

$$|B_{22}(\delta)| \le \delta\, C\, \|\phi\|_{\mathcal{C}^1} \sqrt{C_\Psi\, C_D} \to 0 \text{ as } \delta \to 0.$$

Concerning $B_3(\delta) - B_{30}(\delta)$, we proceed like in the proof of Theorem 3.4.1. We rewrite $B_3(\delta)$ as $B_{31}(\delta) + B_{32}(\delta)$ with

$$B_{31}(\delta) = -\sum_{n=0}^{N_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma\, D\Psi_{K,\sigma}^{n+1}\, \frac{N_L^{n+1} - N_K^{n+1}}{2}\, (\phi_L^n - \phi_K^n)\,,$$

$$B_{32}(\delta) = -\sum_{n=0}^{N_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma\, D\Psi_{K,\sigma}^{n+1}\, N_K^{n+1}\, (\phi_L^n - \phi_K^n)\,.$$

Using the definition of $d\Psi_\delta$ and N_δ, we write $B_{30}(\delta) = B_{310}(\delta) + B_{320}(\delta)$ where

$$B_{310}(\delta) = -\sum_{n=0}^{N_T-1} \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \mathrm{m}(\sigma)\, DN_{K,\sigma}^{n+1}\, D\Psi_{K,\sigma}^{n+1} \frac{1}{\mathrm{m}(\mathrm{T_{K,\sigma}})} \int_{t^n}^{t^{n+1}} \int_{T_{K,\sigma}\cap L} \nabla\phi\cdot\mathbf{n}_{K,\sigma}\, dx\, dt,$$

$$B_{320}(\delta) = -\sum_{n=0}^{N_T-1} \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \mathrm{m}(\sigma)\, N_K^{n+1}\, D\Psi_{K,\sigma}^{n+1} \frac{1}{\mathrm{m}(\mathrm{T_{K,\sigma}})} \int_{t^n}^{t^{n+1}} \int_{T_{K,\sigma}} \nabla\phi\cdot\mathbf{n}_{K,\sigma}\, dx\, dt.$$

Using the regularity of ϕ, there exists $C>0$ which does not depend on δ such that

$$|B_{32}(\delta) - B_{320}(\delta)| \le \delta\, C \sum_{n=0}^{N_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \mathrm{m}(\sigma)\, D_\sigma\Psi^{n+1}\, N_K^{n+1},$$

which finally yields using the discrete L^∞ estimate (2.1.20) on N, the Cauchy-Schwarz inequality and the discrete $L^2(0,T;H^1(\Omega))$ estimate (2.2.11) on Ψ:

$$|B_{32}(\delta) - B_{320}(\delta)| \leq \delta\, C\, M \sqrt{T\,\mathrm{m}(\Omega)}\,\sqrt{C_\Psi} \to 0 \text{ as } \delta \to 0.$$

Moreover, we have

$$\begin{aligned} |B_{31}(\delta)| &\leq \frac{\delta}{2}\|\phi\|_{\mathcal{C}^1} \sum_{n=0}^{N_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma\, D_\sigma \Psi^{n+1}\, D_\sigma N^{n+1} \\ &\leq \frac{\delta}{2}\|\phi\|_{\mathcal{C}^1}\sqrt{C_\Psi\, C_D} \to 0 \text{ as } \delta \to 0. \end{aligned}$$

We obtain in the same way that $B_{310}(\delta) \to 0$ as $\delta \to 0$, which concludes the proof. □

2.4 Numerical experiments

In this section we give some preliminary numerical results concerning the spatial accuracy of hte Scharfetter-Gummel scheme for different values of λ. We consider a one-dimensional test case: Ω is the interval $(0,1)$. The initial and Dirichlet boundary conditions satisfy the quasi-neutrality assumption:

$$P - N + C = 0$$

in order to avoid any initial or boundary layers. We run computations with the fully implicit Scharfetter-Gummel scheme (2.1.13)–(2.1.19) for different doping profiles, with a time step $\Delta t = 10^{-2}$. An estimation of the relative error for the electron density N in L^p norm at time T is given by:

$$e_{2\,\Delta x} = \|N_{\Delta x}(T) - N_{2\,\Delta x}(T)\|_{L^p(\Omega)}\,,$$

where $N_{\Delta x}$ represents the approximation computed from a mesh of size Δx.

Test case 1. We first consider the case of a zero doping profile, since this situation corresponds to the one studied in this chapter. The Dirichlet boundary conditions are given by

$$\begin{cases} N^D(0) = P^D(0) = 1, & \Psi^D(0) = -1, \\ N^D(1) = P^D(1) = 0, & \Psi^D(1) = 1, \end{cases}$$

and the initial conditions are

$$N_0(x) = P_0(x) = \begin{cases} 1 & \text{if} \quad x \leq 0.5, \\ 0 & \text{if} \quad x > 0.5. \end{cases}$$

Test case 2. Then we consider a continuous doping profile: $C(x) = -1$ for $0 \leq x \leq 0.4$, $C(x) = +1$ for $0.6 \leq x \leq 1$ and $C(x)$ is affine on $[0.4; 0.6]$. The initial and boundary conditions are the following:

$$\begin{cases} N^D(0) = 0, & P^D(0) = 1, & \Psi^D(0) = 0, \\ N^D(1) = 1, & P^D(1) = 0, & \Psi^D(1) = 4, \end{cases}$$
$$N_0(x) = \max(C(x), 0), \quad P_0(x) = -\min(C(x), 0), \quad x \in \Omega.$$

Test case 3. We finally change the doping profile for a discontinuous one, which corresponds to the physically relevant hypothesis:

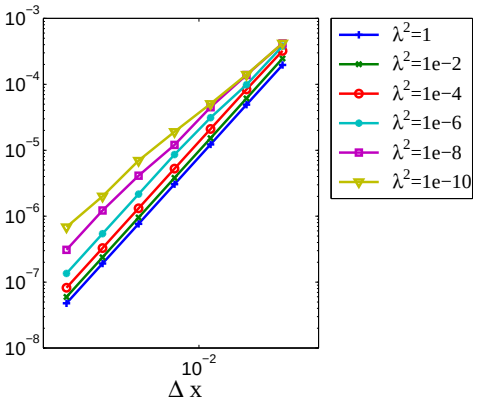
$$C(x) = \begin{cases} -1 & \text{for} \quad x \leq 0.5, \\ +1 & \text{for} \quad x > 0.5. \end{cases}$$

The initial and boundary conditions are the same as in Test 3.

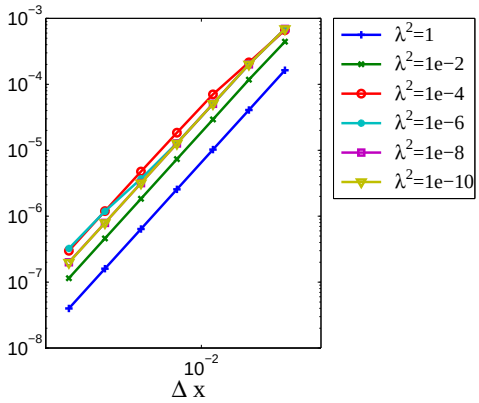
In Figures 2.1 and 2.2, we plot the L^1 and L^2 errors obtained for these three test cases at time $T = 0.1$ for different values of the rescaled Debye length. We first underline the fact that the scheme is unconditionally stable: it is still running even for $\Delta t = 10^2$ and $\lambda^2 = 10^{-10}$. Then, we observe that for a zero or a continuous doping profile, the order of accuracy of the Scharfetter-Gummel scheme (2.1.13)–(2.1.19) is not deteriorated, even for small values of the parameter λ^2. Concerning the test 3, there is a loss of accuracy for values of λ^2 smaller than 10^{-6}, probably due to the discontinuity of the doping profile.

2.5 Conclusion

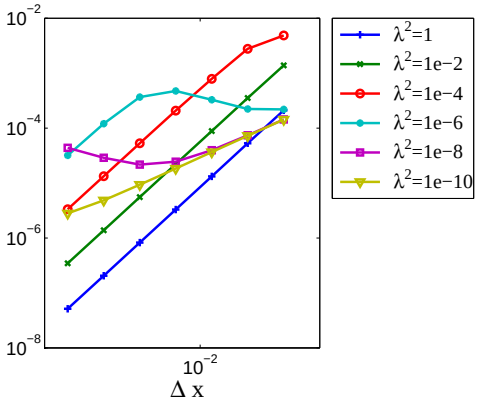
In this chapter, we prove the convergence of the fully implicit Scharfetter-Gummel scheme for all positive values of the rescaled Debye length. This proof is based on a priori estimates which are independent of λ. We also present some preliminary numerical experiments which demonstrate the strong influence of the doping profile regularity. Indeed, the results obtained for a zero or a continuous doping profile are almost the same: the order of accuracy does not depend on the value of λ. On the contrary, we observe a loss of accuracy for small values of λ in the case of a discontinuous doping profile. A future work would be to perform more numerical experiments, including multidimensional and more physically relevant test cases, and to study precisely the assumptions needed on the doping profile to prove the convergence of the scheme independently of λ. Moreover it remains to study the quasi-neutral limit $\lambda \to 0$ in the scheme.



(a) Test 1 ($C = 0$)

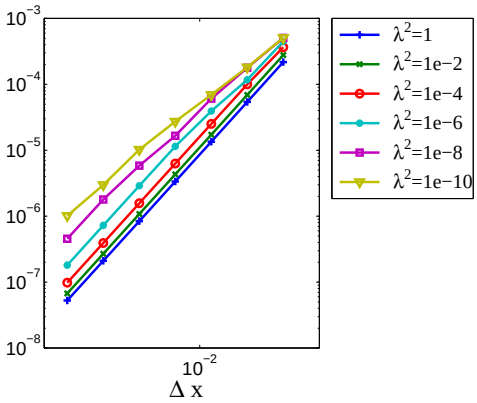


(b) Test 2 (C continuous)

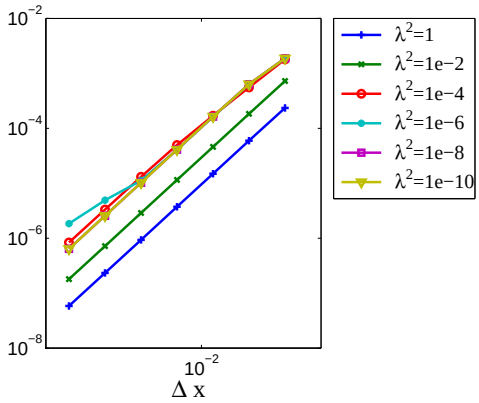


(c) Test 3 (C discontinuous)

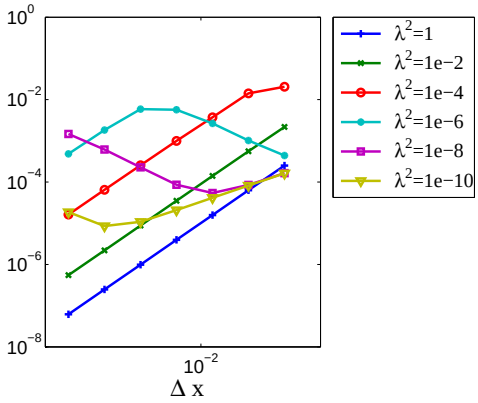
Figure 2.1: L^1 error in log scale at time $T = 0.1$ for different values of λ^2.



(a) Test 1 ($C = 0$)



(b) Test 2 (C continuous)



(c) Test 3 (C discontinuous)

Figure 2.2: L^2 error in log scale at time $T = 0.1$ for different values of λ^2.

Deuxième partie

Schémas volumes finis préservant l'asymptotique en temps long pour des équations paraboliques non linéaires

Dans cette seconde partie, notre objectif est de proposer des schémas volumes finis pour des équations paraboliques non linéaires, éventuellement dégénérées, de la forme suivante :

$$\partial_t u = \operatorname{div}\left(u\,\nabla V(x) + \nabla r(u)\right), \quad x \in \Omega, \quad t > 0, \tag{2.5.1}$$

où $u : \mathbb{R}^+ \times \Omega \to \mathbb{R}^+$ est l'inconnue, V est un potentiel donné et r est une fonction régulière donnée. Au niveau continu, le comportement en temps long de cette équation a été étudié par J. A. Carrillo *et al.* dans [45]. Plus précisément, les auteurs prouvent dans cet article que sous des hypothèses de régularité suffisante sur le potentiel V et la fonction r, la solution converge à une vitesse exponentielle vers un unique état d'équilibre quand $t \to +\infty$. Cette étude repose notamment sur une estimation de l'entropie relative avec contrôle de la dissipation d'entropie. Le but est de proposer et d'étudier des schémas volumes finis permettant d'obtenir un comportement en temps long satisfaisant, avec en particulier un analogue discret de l'estimation d'entropie–dissipation.

En comparant les résultats numériques obtenus avec différents schémas volumes finis, il apparaît fondamental que le flux numérique préserve les états d'équilibre pour obtenir un comportement en temps long cohérent de la solution approchée. Nous constatons notamment que le schéma de Scharfetter-Gummel est très efficace, tant du point de vue de la préservation de l'asymptotique en temps long (voir le chapitre 1) que de la précision (il est d'ordre deux en espace [132]). Néanmoins, il n'est défini que dans le cas d'une diffusion linéaire ($r(u) = u$ dans (2.5.1)). Notre première idée est donc de construire un schéma qui généralise celui de Scharfetter-Gummel pour des équations de convection-diffusion avec une diffusion non linéaire de la forme (2.5.1), en s'attachant à conserver les états stationnaires afin de préserver l'asymptotique en temps long des solutions.

Dans le chapitre 3, nous introduisons et nous étudions une généralisation du schéma de Scharfetter-Gummel, en se plaçant dans le cas non dégénéré (c'est-à-dire $r'(u) \neq 0$). Nous considérons une discrétisation semi-implicite en temps, ce qui permet d'avoir un schéma relativement facile à implémenter, puisqu'il conduit seulement à résoudre un système linéaire à chaque pas de temps. Nous prouvons la compacité d'une famille de solutions approchées par des arguments classiques fondés sur des estimations a priori obtenues à partir du schéma et des estimations des translatées en temps et en espace de la solution approchée. Nous montrons que la limite obtenue est bien solution faible de l'équation (2.5.1), ce qui représente la principale difficulté car les termes de convection et de diffusion sont discrétisés ensembles dans le flux de Scharfetter-Gummel. Nous appliquons finalement ce schéma à deux modèles issus de la physique, l'équation des milieux poreux et le système de dérive-diffusion pour les semi-conducteurs. Nous constatons ainsi que les résultats numériques obtenus sont très satisfaisants si la diffusion ne dégénère pas, aussi bien du point de vue de l'asymptotique en temps long que de l'ordre de convergence en espace, ce qui n'est plus toujours vrai dans le cas dégénéré.

En nous fondant sur cette observation, nous proposons dans le chapitre 4 une nouvelle discrétisation en espace de l'équation (2.5.1). L'idée est toujours de proposer un flux numérique préservant les équilibres, mais qui reste de plus d'ordre élevé même dans le

régime dégénéré. A cette fin, nous prenons en compte ensemble les termes de convection et de diffusion de (2.5.1), en réécrivant le flux sous la forme d'un flux d'advection :

$$u\nabla V + \nabla r(u) = u\,\nabla\left(V + h(u)\right),$$

où la fonction h est telle que $r'(s) = sh'(s)$. Le terme $\nabla\left(V + h(u)\right)$ est alors considéré comme une « vitesse », dans laquelle contribuent à la fois la convection et la diffusion. Nous appliquons ensuite une méthode de discrétisation standard pour les équations de transport : le schéma décentré amont. Ainsi, le flux numérique est défini de telle sorte que les états d'équilibre sont préservés et qu'un analogue discret de l'inégalité d'entropie–dissipation puisse être automatiquement obtenu. Le schéma obtenu est dans ce cas seulement d'ordre 1 en espace. Il suffit ensuite d'appliquer une méthode de limiteurs de pente pour obtenir un schéma qui reste précis à l'ordre 2 en espace, même dans le cas dégénéré.

Ce raisonnement peut se généraliser à d'autres équations admettant une fonctionnelle d'entropie. C'est le cas par exemple pour l'équation suivante, de type Fokker-Planck non linéaire, modélisant les gaz de bosons ($k = 1$) ou de fermions ($k = -1$) :

$$\partial_t u = \operatorname{div}\left(xu(1+ku) + \nabla u\right).$$

Pour définir le schéma numérique pour cette équation, nous réécrivons le flux de la manière suivante :

$$xu(1+ku) + \nabla u = u(1+ku)\nabla\left(\frac{|x|^2}{2} + \log\left(\frac{u}{1+ku}\right)\right)$$

et nous appliquons un schéma de Lax-Friedrichs local pour définir une approximation de ce flux à l'interface ; nous pouvons alors utiliser une méthode de limiteurs de pente pour obtenir un schéma d'ordre élevé.

L'étude d'un schéma d'ordre élevé est relativement délicate. Nous pouvons prouver dans notre cas la positivité de la solution numérique obtenue, ainsi qu'une estimation d'entropie semi-discrète en espace. Nous vérifions numériquement que le schéma reste précis à l'ordre 2 en espace même dans le cas dégénéré, et nous constatons son efficacité pour préserver l'asymptotique en temps long en l'appliquant à différents modèles issus de la physique : équation des milieux poreux, équation de Fokker-Planck non linéaire pour les bosons et les fermions, système de dérive-diffusion pour les semi-conducteurs, équation de Buckley-Leverett.

Chapitre 3

Un schéma volumes finis pour des équations de convection-diffusion avec diffusion non linéaire dérivé du schéma de Scharfetter-Gummel *

Dans ce chapitre, nous proposons et étudions un schéma volumes finis pour des équations de convection-diffusion dont le terme de diffusion est non linéaire. De telles équations apparaissent dans de nombreux contextes physiques. En particulier, nous nous concentrons sur le système de dérive-diffusion pour les semi-conducteurs et sur l'équation des milieux poreux. Dans ces deux cas, il a été démontré que la solution du problème évolutif converge vers une solution stationnaire quand t tend vers l'infini.
Le schéma proposé est une généralisation du schéma de Scharfetter-Gummel pour une diffusion non linéaire. Il reste valable même quand l'équation dégénère, et il préserve les états stationnaires. Nous prouvons la convergence de ce schéma dans le cas non dégénéré. Finalement, nous présentons des simulations numériques appliquées aux deux modèles physiques introduits, et nous soulignons l'efficacité du schéma pour préserver le comportement en temps long des solutions.

∗. Ce chapitre est un article accepté pour publication dans Numerische Mathematik, *A finite volume scheme for convection-diffusion equations with nonlinear diffusion derived from the Scharfetter-Gummel scheme* [16].

Contents

3.1 Introduction

In this chapter, our aim is to elaborate a finite volume scheme for convection-diffusion equations with nonlinear diffusion. The main objective of building such a scheme is to preserve steady-states in order to be able to apply it to physical models in which it has been proved that the solution converges to equilibrium in long time. In particular, this convergence can be observed in the drift-diffusion system for semiconductors as well as in the porous media equation.
In this context, we will first present these two physical models – drift-diffusion system for semiconductors and porous media equation. Then, we will precise the general framework of our study in this chapter.

3.1.1 The drift-diffusion model for semiconductors

The drift-diffusion system consists of two continuity equations for the electron density $N(x,t)$ and the hole density $P(x,t)$, as well as a Poisson equation for the electrostatic potential $V(x,t)$, for $t \in \mathbb{R}^+$ and $x \in \mathbb{R}^d$.
Let $\Omega \subset \mathbb{R}^d$ $(d \geq 1)$ be an open and bounded domain. The drift-diffusion system reads

$$\left\{\begin{array}{lll} \partial_t N - \operatorname{div}(\nabla r(N) - N\nabla V) = 0 & \text{on} & \Omega \times (0,T), \\ \partial_t P - \operatorname{div}(\nabla r(P) + P\nabla V) = 0 & \text{on} & \Omega \times (0,T), \\ \Delta V = N - P - C & \text{on} & \Omega \times (0,T), \end{array}\right. \tag{3.1.1}$$

where $C \in L^\infty(\Omega)$ is the prescribed doping profile.
The pressure has the form of a power law,

$$r(s) = s^\gamma, \quad \gamma \geq 1.$$

We supplement these equations with initial conditions $N_0(x)$ and $P_0(x)$ and physically motivated boundary conditions: the boundary $\Gamma = \partial\Omega$ is split into two parts $\Gamma = \Gamma^D \cup \Gamma^N$ and the boundary conditions are Dirichlet boundary conditions $\overline{N}$, $\overline{P}$ and $\overline{V}$ on ohmic contacts Γ^D and homogeneous Neumann boundary conditions on $r(N)$, $r(P)$ and V on insulating boundary segments Γ^N.
The large time behavior of the solutions to the nonlinear drift-diffusion model (3.1.1) has been studied by A. Jüngel in [119]. It is proved that the solution to the transient system converges to a solution of the thermal equilibrium state as $t \to \infty$ if the Dirichlet boundary conditions are in thermal equilibrium. The thermal equilibrium is a particular steady-state for which electron and hole currents, namely $\nabla r(N) - N\nabla V$ and $\nabla r(P) + P\nabla V$, vanish. The existence of a thermal equilibrium has been studied in the case of a linear pressure by P. Markowich, C. Ringhofer and C. Schmeiser in [136, 137], and in the nonlinear case by P. Markowich and A. Unterreiter in [138].
We introduce the enthalpy function h defined by

$$h(s) = \int_1^s \frac{r'(\tau)}{\tau}\, d\tau \tag{3.1.2}$$

and the generalized inverse g of h defined by

$$g(s) = \begin{cases} h^{-1}(s) & \text{if} \quad h(0^+) < s < \infty, \\ 0 & \text{if} \quad s \leq h(0^+). \end{cases}$$

If the boundary conditions satisfy $\overline{N}, \overline{P} > 0$ and

$$h(\overline{N}) - \overline{V} = \alpha_N \text{ and } h(\overline{P}) + \overline{V} = \alpha_P \text{ on } \Gamma^D,$$

the thermal equilibrium is defined by

$$N^{eq}(x) = g\left(\alpha_N + V^{eq}(x)\right), \quad P^{eq}(x) = g\left(\alpha_P - V^{eq}(x)\right), \quad x \in \Omega, \tag{3.1.3}$$

while V^{eq} satisfies the following elliptic problem

$$\begin{cases} \Delta V^{eq} = g\left(\alpha_N + V^{eq}\right) - g\left(\alpha_P - V^{eq}\right) - C \text{ in } \Omega, \\ \\ V^{eq}(x) = \overline{V}(x) \text{ on } \Gamma^D, \quad \nabla V^{eq} \cdot \mathbf{n} = 0 \text{ on } \Gamma^N. \end{cases} \tag{3.1.4}$$

The proof of the convergence to thermal equilibrium is based on an energy estimate with the control of the energy dissipation. More precisely, if we define

$$H(s) = \int_1^s h(\tau) d\tau, \quad s \geq 0, \tag{3.1.5}$$

then we can introduce the deviation of the total energy (sum of the internal energies for the electron and hole densities and the energy due to the electrostatic potential) from the thermal equilibrium (see [119])

$$\begin{aligned}\mathcal{E}(t) &= \int_\Omega \Big(H\left(N(t)\right) - H\left(N^{eq}\right) - h\left(N^{eq}\right)\left(N(t) - N^{eq}\right) + H\left(P(t)\right) - H\left(P^{eq}\right) \\ &\quad -h\left(P^{eq}\right)\left(P(t) - P^{eq}\right) + \frac{1}{2}\left|\nabla\left(V(t) - V^{eq}\right)\right|^2 \Big)\, dx, \end{aligned} \tag{3.1.6}$$

and the energy dissipation

$$\mathcal{I}(t) = \int_\Omega \left(N(t)\left|\nabla(h(N(t)) - V(t))\right|^2 + P(t)\left|\nabla(h(P(t)) + V(t))\right|^2 \right) dx. \tag{3.1.7}$$

Then the keypoint of the proof is the following estimate:

$$0 \le \mathcal{E}(t) + \int_0^t \mathcal{I}(\tau)\, d\tau \le \mathcal{E}(0). \tag{3.1.8}$$

3.1.2 The porous media equation

The flow of a gas in a d-dimensional porous medium is classically described by the Leibenzon-Muskat model,

$$\begin{cases} \partial_t v = \Delta v^\gamma & \text{on } \mathbb{R}^d \times (0,T), \\ v(x,0) = v_0(x) & \text{on } \mathbb{R}^d, \end{cases} \tag{3.1.9}$$

where the function v represents the density of the gas in the porous medium and $\gamma > 1$ is a physical constant.
With a time-dependent scaling (see [48]), we transform (3.1.9) into the nonlinear Fokker-Planck equation

$$\begin{cases} \partial_t u = \operatorname{div}(xu + \nabla u^\gamma) & \text{on } \mathbb{R}^d \times (0,T), \\ u(x,0) = u_0(x) & \text{on } \mathbb{R}^d. \end{cases} \tag{3.1.10}$$

It is proved in [48] that the unique stationary solution of (3.1.10) is given by the Barenblatt-Pattle type formula

$$u^{eq}(x) = \left(C_1 - \frac{\gamma - 1}{2\gamma}|x|^2 \right)_+^{1/(\gamma-1)}, \tag{3.1.11}$$

where C_1 is a constant such that u^{eq} has the same mass as the initial data u_0.
Moreover, J. A. Carrillo and G. Toscani have proved in [48] the convergence of the solution $u(x,t)$ of (3.1.9) to the Barenblatt-Pattle solution $u^{eq}(x)$ as $t \to \infty$. As in the case of the drift-diffusion model, the proof of the convergence to the Barenblatt-Pattle solution is based on an entropy estimate with the control of the entropy dissipation given by (3.1.8), where the relative entropy is defined by

$$\mathcal{E}(t) = \int_{\mathbb{R}^d} \left(H(u(t)) - H(u^{eq}) + \frac{|x|^2}{2}\left(u(t) - u^{eq}\right) \right) dx, \tag{3.1.12}$$

where H is defined by (3.1.5) and the entropy dissipation is given by

$$\mathcal{I}(t) = -\frac{d}{dt}\mathcal{E}(t) = \int_{\mathbb{R}^d} u(t) \left| \nabla \left(h(u(t)) + \frac{|x|^2}{2} \right) \right|^2 dx. \tag{3.1.13}$$

3.1.3 Motivation

Many numerical schemes have been proposed to approximate the solutions of nonlinear convection-diffusion equations. In particular, finite volume methods have been proved to be efficient in the case of degenerate parabolic equations (see [85, 87]). We also mention the combined finite volume-finite element approach for nonlinear degenerate parabolic convection-diffusion-reaction equations analysed in [88]. The definition of the so-called local Péclet upstream weighting numerical flux guarantees the stability of the scheme while reducing the excessive numerical diffusion added by the classical upwinding.
On the other hand, there exists a wide literature on numerical schemes for the drift-diffusion equations. It started with 1-D finite difference methods and the Scharfetter-Gummel scheme ([158]). In the linear pressure case ($r(s) = s$), a mixed exponential fitting finite element scheme has been successfully developed by F. Brezzi, L. Marini and P. Pietra in [33, 34]. The adaptation of the mixed exponential fitting method to the nonlinear case has been developed by F. Arimburgo, C. Baiocchi, L. Marini in [9] and by A. Jüngel in [118] for the one-dimensional problem, and by A. Jüngel and P. Pietra in [122] for the two-dimensional problem. Moreover, C. Chainais-Hillairet and Y. J. Peng proposed a finite volume scheme for the drift-diffusion equations in 1-D in [53], which was extended in [52, 54] in the multidimensional case. C. Chainais-Hillairet and F. Filbet also introduced in [51] a finite-volume scheme preserving the large time behavior of the solutions of the nonlinear drift-diffusion model.
Now to explain our approach, let us first recall some previous numerical results concerning the drift-diffusion system for semiconductors. The precise definitions of schemes considered will be presented in Section 3.2. We compare results obtained with three existing finite volume schemes: the classical upwind scheme proposed by C. Chainais-Hillairet and Y. J. Peng in [53], the Scharfetter-Gummel scheme introduced in [158] and the nonlinear upwind scheme studied in [51].
In Figure 3.1, we present some results obtained in the case of a linear diffusion ($r(s) = s$). We represent the relative energy $\mathcal{E}$ and the dissipation of energy $\mathcal{I}$ obtained with the upwind flux and the Scharfetter-Gummel flux for a test case in one space dimension. We can observe a phenomenon of saturation of $\mathcal{E}$ and $\mathcal{I}$ for the upwind flux. In addition, we clearly observe that the energy and its dissipation obtained with the Scharfetter-Gummel flux converge to zero when time goes to infinity, which means that densities $N(t)$ and $P(t)$ converge to the thermal equilibrium. It appears that the Scharfetter-Gummel flux is very efficient, but is only valid for linear diffusion. Moreover, we can emphasize that contrary to the upwind flux, the Scharfetter-Gummel flux preserves the thermal equilibrium.
In Figure 3.2, we present numerical results obtained in the case of a nonlinear diffusion $r(s) = s^2$. We represent the relative energy $\mathcal{E}$ and the dissipation $\mathcal{I}$ obtained with the classical upwind flux and with the nonlinear upwind flux for a test case in one dimension

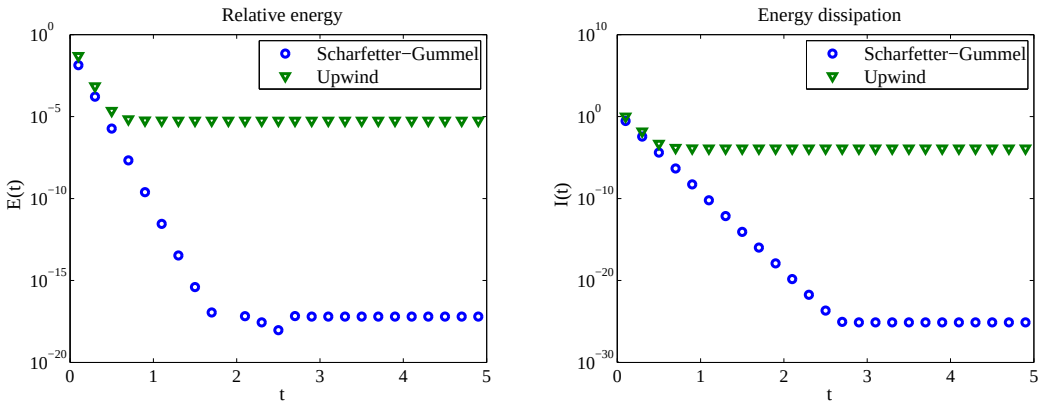


Figure 3.1: Linear case: relative energy $\mathcal{E}^n$ and dissipation $\mathcal{I}^n$ for different schemes in log scale, with time step $\Delta t = 10^{-2}$ and space step $\Delta x = 10^{-2}$.

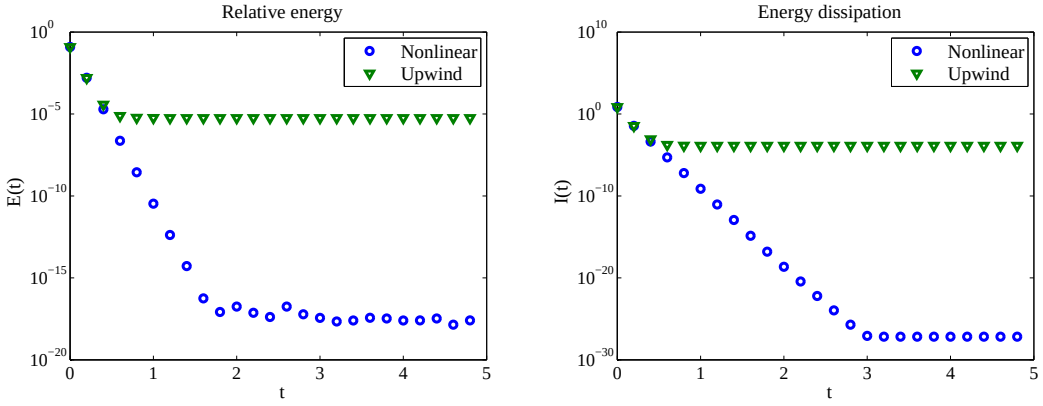


Figure 3.2: Nonlinear case: relative energy $\mathcal{E}^n$ and dissipation $\mathcal{I}^n$ for different schemes in log scale, with time step $\Delta t = 5.10^{-4}$ and space step $\Delta x = 10^{-2}$.

of space. We still observe a phenomenon of saturation of $\mathcal{E}$ and $\mathcal{I}$ for the classical upwind flux. For the nonlinear flux, we clearly notice that the energy and its dissipation converge to zero when time goes to infinity.
Looking at these results, it seems crucial that the numerical flux preserves the thermal equilibrium to obtain the consistency of the approximate solution in the long time asymptotic limit.

Our aim is to propose a finite volume scheme for convection-diffusion equations with nonlinear diffusion. We will focus on preserving steady-states in order to obtain a satisfying long-time behavior of the approximate solution. The scheme proposed in [51] satisfies this property but because of the nonlinear discretization of the diffusive terms, it leads to solve a nonlinear system at each time step, even in the case of a linear diffusion. The idea is to extend the Scharfetter-Gummel scheme, which is only valid in the case of a

linear diffusion, for convection-diffusion equations with nonlinear diffusion, even in the degenerate case. Some extensions of this scheme have already been proposed. Indeed, R. Eymard, J. Fuhrmann and K. Gärtner studied a scheme valid in the case where the convection and diffusion terms are nonlinear (see [83]), but their method leads to solve a nonlinear elliptic problem at each interface. A. Jüngel and P. Pietra proposed a scheme for the drift-diffusion model (see [118, 122]), but it is not very satisfying to reflect the large-time behavior of the solutions.

3.1.4 General framework

We will now consider the following problem:

$$\partial_t u - \operatorname{div}(\nabla r(u) - \mathbf{q}u) = 0 \text{ for } (x,t) \in \Omega \times (0,T), \tag{3.1.14}$$

with an initial condition

$$u(x,0) = u_0(x) \text{ for } x \in \Omega. \tag{3.1.15}$$

Moreover, we will consider Dirichlet-Neumann boundary conditions. The boundary $\partial\Omega = \Gamma$ is split into two parts $\Gamma = \Gamma^D \cup \Gamma^N$ and, if we denote by $\mathbf{n}$ the outward normal to Γ, the boundary conditions are Dirichlet boundary conditions on Γ^D

$$u(x,t) = \overline{u}(x,t) \text{ for } (x,t) \in \Gamma^D \times (0,T), \tag{3.1.16}$$

and homogeneous Neumann boundary conditions on Γ^N:

$$\nabla r(u) \cdot \mathbf{n} = 0 \text{ on } \Gamma^N \times (0,T). \tag{3.1.17}$$

Remark 1. We will construct the scheme and perform some numerical experiments in the case of Dirichlet-Neumann boundary conditions. However, for the analysis of the scheme, we will only consider the case of Dirichlet boundary conditions ($\partial\Omega = \Gamma^D = \Gamma$).

We suppose that the following hypotheses are fulfilled:

(H1) Ω is an open bounded connected subset of $\mathbb{R}^d$, with $d = 1, 2$ or 3,

(H2) $\partial\Omega = \Gamma^D = \Gamma$, $\overline{u}$ is the trace on $\Gamma \times (0,T)$ of a function, also denoted $\overline{u}$, which is assumed to satisfy $\overline{u} \in H^1(\Omega \times (0,T)) \cap L^\infty(\Omega \times (0,T))$ and $\overline{u} \geq 0$ *a.e.*,

(H3) $u_0 \in L^\infty(\Omega)$ and $u_0 \geq 0$ *a.e.*,

(H4) $r \in C^2(\mathbb{R})$ is strictly increasing on $]0, +\infty[$, $r(0) = r'(0) = 0$, with $r'(s) \geq c_0 s^{\gamma-1}$,

(H5) $\mathbf{q} \in C^1(\overline{\Omega}, \mathbb{R}^d)$.

H. Alt, S. Luckhaus and A. Visintin, as well as J. Carrillo, studied the existence and uniqueness of a weak solution to the problem (3.1.14)-(3.1.17) in [2] and [42] respectively.

Definition 3.1. We say that u is a solution to the problem (3.1.14)-(3.1.15)-(3.1.16)-(3.1.17) if it verifies:

$$u \in L^\infty(\Omega \times (0,T)), \quad u - \overline{u} \in L^2(0,T; H_0^1(\Omega))$$

and for all $\psi \in \mathcal{D}(\Omega \times [0,T[)$,

$$\int_0^T \int_\Omega (u\,\partial_t\psi - \nabla(r(u)) \cdot \nabla\psi + u\,\mathbf{q} \cdot \nabla\psi)\,dx\,dt + \int_\Omega u(x,0)\,\psi(x,0)\,dx = 0. \tag{3.1.18}$$

The outline of the chapter is the following. In Section 3.2, we construct the finite volume scheme. In Section 3.3, we prove the existence and uniqueness of the solution of the scheme and give some estimates on this solution. Then, thanks to these estimates, we prove in Section 3.4 the compactness of a family of approximate solutions. It yields the convergence (up to a subsequence) of the solution u_δ of the scheme to a solution of (3.1.14)-(3.1.17) when δ goes to 0. In the last section, we present some numerical results that show the efficiency of the scheme.

3.2 Presentation of the numerical scheme

In this section, we present our new finite volume scheme for equation (3.1.14) and other existing schemes. We will then compare these schemes to our new one.

3.2.1 Definition of the finite volume scheme

We first define the space discretization of Ω. A regular and admissible mesh of Ω is given by a family $\mathcal{T}$ of control volumes (open and convex polygons in 2-D, polyhedra in 3-D), a family $\mathcal{E}$ of edges in 2-D (faces in 3-D) and a family of points $(x_K)_{K\in\mathcal{T}}$ which satisfy Definition 5.1 in [85]. It implies that the straight line between two neighboring centers of cells (x_K, x_L) is orthogonal to the edge $\sigma = K|L$.
In the set of edges $\mathcal{E}$, we distinguish the interior edges $\sigma \in \mathcal{E}_{int}$ and the boundary edges $\sigma \in \mathcal{E}_{ext}$. Because of the Dirichlet-Neumann boundary conditions, we split $\mathcal{E}_{ext}$ into $\mathcal{E}_{ext} = \mathcal{E}_{ext}^D \cup \mathcal{E}_{ext}^N$ where $\mathcal{E}_{ext}^D$ is the set of Dirichlet boundary edges and $\mathcal{E}_{ext}^N$ is the set of Neumann boundary edges. For a control volume $K \in \mathcal{T}$, we denote by $\mathcal{E}_K$ the set of its edges, $\mathcal{E}_{int,K}$ the set of its interior edges, $\mathcal{E}_{ext,K}^D$ the set of edges of K included in Γ^D and $\mathcal{E}_{ext,K}^N$ the set of edges of K included in Γ^N.
The size of the mesh is defined by

$$\Delta x = \max_{K\in\mathcal{T}}(\mathrm{diam}(K)).$$

In the sequel, we denote by d the distance in $\mathbb{R}^d$ and m the measure in $\mathbb{R}^d$ or $\mathbb{R}^{d-1}$.
We note for all $\sigma \in \mathcal{E}$

$$d_\sigma = \begin{cases} \mathrm{d}(x_K, x_L), & \text{for } \sigma \in \mathcal{E}_{int}, \qquad \sigma = K|L, \\ \mathrm{d}(x_K, \sigma), & \text{for } \sigma \in \mathcal{E}_{ext,K}. \end{cases}$$

For all $\sigma \in \mathcal{E}$, we define the transmissibility coefficient $\tau_\sigma = \mathrm{m}(\sigma)/d_\sigma$. For $\sigma \in \mathcal{E}_K$, $\mathbf{n}_{K,\sigma}$ is the unit vector normal to σ outward to K.
We may now define the finite volume approximation of the equation (3.1.14)-(3.1.17).
Let $(\mathcal{T}, \mathcal{E}, (x_K)_{K\in\mathcal{T}})$ be an admissible discretization of Ω and let us define the time step Δt, $N_T = E(T/\Delta t)$ and the increasing sequence $(t^n)_{0\le n\le N_T}$, where $t^n = n\Delta t$, in order to get a space-time discretization $\mathcal{D}$ of $\Omega \times (0, T)$. The size of the space-time discretization $\mathcal{D}$ is defined by:

$$\delta = \max(\Delta x, \Delta t).$$

First of all, the initial condition is discretized by:

$$U_K^0 = \frac{1}{\mathrm{m}(K)} \int_K u_0(x)\, dx, \quad K \in \mathcal{T}. \tag{3.2.1}$$

In order to introduce the finite volume scheme, we also need to define the numerical boundary conditions:

$$U_\sigma^{n+1} = \frac{1}{\Delta t\, \mathrm{m}(\sigma)} \int_{t^n}^{t^{n+1}} \int_\sigma \overline{u}(s,t)\, ds\, dt, \quad \sigma \in \mathcal{E}_{ext}^D, \ \ n \geq 0. \tag{3.2.2}$$

We set

$$q_{K,\sigma} = \frac{1}{\mathrm{m}(\sigma)} \int_\sigma \mathbf{q}(x) \cdot \mathbf{n}_{K,\sigma}\, ds(x), \quad \forall K \in \mathcal{T}, \quad \forall \sigma \in \mathcal{E}_K. \tag{3.2.3}$$

The finite volume scheme is obtained by integrating the equation (3.1.14) on each control volume and by using the divergence theorem. We choose a backward Euler discretization in time (in order to avoid a restriction on the time step of the form $\Delta t = O(\Delta x^2)$). Then the scheme on u is given by the following set of equations:

$$\mathrm{m}(K) \frac{U_K^{n+1} - U_K^n}{\Delta t} + \sum_{\sigma \in \mathcal{E}_K} \mathcal{F}_{K,\sigma}^{n+1} = 0, \tag{3.2.4}$$

where the numerical flux $\mathcal{F}_{K,\sigma}^{n+1}$ is an approximation of $-\displaystyle\int_\sigma (\nabla r(u) - \mathbf{q}u) \cdot \mathbf{n}_{K,\sigma}$ which remains to be defined.

3.2.2 Definition of the numerical flux

Existing schemes

We presented in introduction some numerical results obtained with different choices of numerical fluxes for the drift-diffusion system. We are now going to define precisely these fluxes.

The classical upwind flux. This flux was studied in [85] for a scalar convection-diffusion equation. It is valid both in the case of a linear diffusion and in the case of a nonlinear diffusion. The diffusion term is discretized classically by using a two-points flux and the convection term is discretized with the upwind flux, whose origin can be traced back to the work of R. Courant, E. Isaacson and M. Rees [64]. This flux was then used for the drift-diffusion system for semiconductors in [53] and [52, 54] in 1-D and in 2-D respectively. The definition of this flux is

$$\mathcal{F}_{K,\sigma}^{n+1} = \begin{cases} \tau_\sigma \left(r\left(U_K^{n+1}\right) - r\left(U_L^{n+1}\right) + d_\sigma \left(q_{K,\sigma}^+ U_K^{n+1} - q_{K,\sigma}^- U_L^{n+1} \right) \right), & \forall \sigma = K|L, \\ \tau_\sigma \left(r\left(U_K^{n+1}\right) - r\left(U_\sigma^{n+1}\right) + d_\sigma \left(q_{K,\sigma}^+ U_K^{n+1} - q_{K,\sigma}^- U_\sigma^{n+1} \right) \right), & \forall \sigma \in \mathcal{E}_{ext,K}^D, \\ 0, & \forall \sigma \in \mathcal{E}_{ext,K}^N, \end{cases} \tag{3.2.5}$$

where $s^+ = \max(s,0)$ and $s^- = \max(-s,0)$ are the positive and negative parts of a real number s.

The upwind flux with nonlinear discretization of the diffusion term. This flux was introduced in [51] in the context of the drift-diffusion system for semiconductors. The idea is to write the flux $-\int_\sigma (\nabla r(u) - \mathbf{q}u)\cdot \mathbf{n}_{K,\sigma}$ as $-\int_\sigma (u\nabla h(u) - \mathbf{q}u)\cdot \mathbf{n}_{K,\sigma}$, where h is the enthalpy function defined by (3.1.2). The flux is then defined with a standard upwinding for the convective term and a nonlinear approximation for the diffusive term:

$$\mathcal{F}_{K,\sigma}^{n+1} = \begin{cases} -\tau_\sigma\Big(\min\left(U_K^{n+1}, U_L^{n+1}\right) Dh\left(U^{n+1}\right)_{K,\sigma} + d_\sigma\left(q_{K,\sigma}^+ U_K^{n+1} - q_{K,\sigma}^- U_L^{n+1}\right)\Big), & \forall \sigma = K|L, \\ -\tau_\sigma\Big(\min\left(U_K^{n+1}, U_\sigma^{n+1}\right) Dh\left(U^{n+1}\right)_{K,\sigma} + d_\sigma\left(q_{K,\sigma}^+ U_K^{n+1} - q_{K,\sigma}^- U_\sigma^{n+1}\right)\Big), & \forall \sigma \in \mathcal{E}_{ext,K}^D, \\ 0, & \forall \sigma \in \mathcal{E}_{ext,K}^N, \end{cases}$$

where for a given function f, $Df(U)_{K,\sigma}$ is defined by

$$Df(U)_{K,\sigma} = \begin{cases} f(U_L) - f(U_K), & \text{if } \sigma = K|L \in \mathcal{E}_{K,int}, \\ f(U_\sigma) - f(U_K), & \text{if } \sigma \in \mathcal{E}_{K,ext}^D, \\ 0, & \text{if } \sigma \in \mathcal{E}_{K,ext}^N. \end{cases}$$

This flux preserves the thermal equilibrium and it is proved that the numerical solution converges to this equilibrium when time goes to infinity.

The Scharfetter-Gummel flux. This flux is widely used in the semiconductors framework in the case of a linear diffusion, namely $r(s) = s$. It has been proposed by D. L. Scharfetter and H. K. Gummel in [158] for the numerical approximation of the one-dimensional drift-diffusion model. We also refer to the work of A. M. Il'in [114], where the same kind of flux was introduced for one-dimensional finite-difference schemes. The Scharfetter-Gummel flux preserves steady-state, and is second order accurate in space (see [132]). It is defined by:

$$\mathcal{F}_{K,\sigma}^{n+1} = \begin{cases} \tau_\sigma\left(B(-d_\sigma q_{K,\sigma}) U_K^{n+1} - B(d_\sigma q_{K,\sigma}) U_L^{n+1}\right), & \forall \sigma = K|L \in \mathcal{E}_{K,int}, \\ \tau_\sigma\left(B(-d_\sigma q_{K,\sigma}) U_K^{n+1} - B(d_\sigma q_{K,\sigma}) U_\sigma^{n+1}\right), & \forall \sigma \in \mathcal{E}_{K,ext}^D, \\ 0, & \forall \sigma \in \mathcal{E}_{K,ext}^N, \end{cases}$$

where B is the Bernoulli function defined by

$$B(x) = \frac{x}{e^x - 1} \text{ for } x \neq 0, \quad B(0) = 1.$$

Extension of the Scharfetter-Gummel flux

Now we will extend the Scharfetter-Gummel flux to the case of a nonlinear diffusion. Firstly, if we consider the linear case with a viscosity coefficient $\varepsilon > 0$, namely

$$\partial_t u - \operatorname{div}(\varepsilon \nabla u - \mathbf{q}u) = 0 \text{ for } (x,t) \in \Omega \times (0,T),$$

then the Scharfetter-Gummel flux is defined by:

$$\mathcal{F}_{K,\sigma}^{n+1} = \tau_\sigma \varepsilon \left(B\left(\frac{-d_\sigma q_{K,\sigma}}{\varepsilon}\right) U_K^{n+1} - B\left(\frac{d_\sigma q_{K,\sigma}}{\varepsilon}\right) U_L^{n+1} \right) \quad \forall \sigma = K|L \in \mathcal{E}_{int,K}. \tag{3.2.6}$$

Using the following properties of the Bernoulli function:

$$B(s) \underset{s\to+\infty}{\longrightarrow} 0 \text{ and } B(s) \underset{-\infty}{\sim} -s,$$

it is clear that if ε tends to zero, this flux degenerates into the classical upwind flux for the transport equation $\partial_t u - \operatorname{div}(\mathbf{q}u) = 0$:

$$\mathcal{F}_{K,\sigma}^{n+1} = \mathrm{m}(\sigma) \left(q_{K,\sigma}^+ U_K^{n+1} - q_{K,\sigma}^- U_L^{n+1} \right) \quad \forall \sigma = K|L \in \mathcal{E}_{int,K}. \tag{3.2.7}$$

Now considering a nonlinear diffusion, we can write $\nabla r(u)$ as $r'(u)\nabla u$. We denote by $dr_{K,\sigma}$ an approximation of $r'(u)$ at the interface $\sigma \in \mathcal{E}_K$, which will be defined later. We consider this term as a viscosity coefficient and then, using (3.2.6), we extend the Scharfetter-Gummel flux by defining:

$$\mathcal{F}_{K,\sigma}^{n+1} = \begin{cases} \tau_\sigma dr_{K,\sigma} \left(B\left(\dfrac{-d_\sigma q_{K,\sigma}}{dr_{K,\sigma}}\right) U_K^{n+1} - B\left(\dfrac{d_\sigma q_{K,\sigma}}{dr_{K,\sigma}}\right) U_L^{n+1} \right), & \forall \sigma = K|L, \\ \tau_\sigma dr_{K,\sigma} \left(B\left(\dfrac{-d_\sigma q_{K,\sigma}}{dr_{K,\sigma}}\right) U_K^{n+1} - B\left(\dfrac{d_\sigma q_{K,\sigma}}{dr_{K,\sigma}}\right) U_\sigma^{n+1} \right), & \forall \sigma \in \mathcal{E}_{ext,K}^D, \\ 0, & \forall \sigma \in \mathcal{E}_{ext,K}^N. \end{cases} \tag{3.2.8}$$

In the degenerate case, $dr_{K,\sigma}$ can vanish and then this flux degenerates into the upwind flux (3.2.7). Now it remains to define $dr_{K,\sigma}$.

Definition of $dr_{K,\sigma}$. A first possibility is to take the value of r' at the average of U_K and U_σ:

$$dr_{K,\sigma} = \begin{cases} r'\left(\dfrac{U_K + U_L}{2}\right), & \forall \sigma = K|L \in \mathcal{E}_{int,K}, \\ r'\left(\dfrac{U_K + U_\sigma}{2}\right), & \forall \sigma \in \mathcal{E}_{ext,K}^D. \end{cases} \tag{3.2.9}$$

This choice is quite close to the one of A. Jüngel and P. Pietra (see [118, 122]). However, considering the numerical results presented in the introduction, it seems important that

the numerical flux preserves the equilibrium. Therefore, we define the function dr as follows: for a, $b \in \mathbb{R}_+$,

$$dr(a,b) = \begin{cases} \dfrac{h(b)-h(a)}{\log(b)-\log(a)} & \text{if } ab>0 \text{ and } a \neq b, \\ \\ r'\left(\dfrac{a+b}{2}\right) & \text{elsewhere,} \end{cases} \tag{3.2.10}$$

and we set for all $K \in \mathcal{T}$

$$dr_{K,\sigma} = \begin{cases} dr(U_K, U_L), & \text{for } \sigma = K|L \in \mathcal{E}_{K,int}, \\ dr(U_K, U_\sigma), & \text{for } \sigma \in \mathcal{E}_{K,ext}^D. \end{cases} \tag{3.2.11}$$

Remark 2. Let $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$. We assume that $dr_{K,\sigma}$ is defined by (3.2.11) in (3.2.8) and that $U_K > 0$ and $U_\sigma > 0$. If $d_\sigma q_{K,\sigma} = Dh(U)_{K,\sigma}$, then $\mathcal{F}_{K,\sigma} = 0$.
Indeed,

$$\begin{aligned} \mathcal{F}_{K,\sigma} &= \tau_\sigma dr_{K,\sigma}\left(B\left(-\frac{Dh(U)_{K,\sigma}}{dr_{K,\sigma}}\right)U_K - B\left(\frac{Dh(U)_{K,\sigma}}{dr_{K,\sigma}}\right)U_\sigma\right) \\ &= \tau_\sigma Dh(U)_{K,\sigma}\left(\frac{\exp\left(\dfrac{Dh(U)_{K,\sigma}}{dr_{K,\sigma}}\right)U_K - U_\sigma}{\exp\left(\dfrac{Dh(U)_{K,\sigma}}{dr_{K,\sigma}}\right)-1}\right). \end{aligned}$$

But using the definition (3.2.10) of dr, we obtain

$$\exp\left(\frac{Dh(U)_{K,\sigma}}{dr_{K,\sigma}}\right) = \frac{U_\sigma}{U_K},$$

and then $\mathcal{F}_{K,\sigma} = 0$. Thus the scheme preserves this type of steady-state.

Time discretization. We choose an explicit expression of $dr_{K,\sigma}$:

$$dr_{K,\sigma}^n = \begin{cases} dr(U_K^n, U_L^n), & \text{for } \sigma = K|L \in \mathcal{E}_{K,int}, \\ dr(U_K^n, U_\sigma^n), & \text{for } \sigma \in \mathcal{E}_{K,ext}^D. \end{cases} \tag{3.2.12}$$

Thus we obtain a scheme which leads only to solve a linear system of equations at each time step.
To sum up, our extension of the Scharfetter-Gummel flux is defined by

$$\mathcal{F}_{K,\sigma}^{n+1} = \begin{cases} \tau_\sigma dr_{K,\sigma}^n\left(B\left(\dfrac{-d_\sigma q_{K,\sigma}}{dr_{K,\sigma}^n}\right)U_K^{n+1} - B\left(\dfrac{d_\sigma q_{K,\sigma}}{dr_{K,\sigma}^n}\right)U_L^{n+1}\right), & \forall \sigma = K|L, \\ \\ \tau_\sigma dr_{K,\sigma}^n\left(B\left(\dfrac{-d_\sigma q_{K,\sigma}}{dr_{K,\sigma}^n}\right)U_K^{n+1} - B\left(\dfrac{d_\sigma q_{K,\sigma}}{dr_{K,\sigma}^n}\right)U_\sigma^{n+1}\right), & \forall \sigma \in \mathcal{E}_{K,ext}^D, \\ \\ 0, & \forall \sigma \in \mathcal{E}_{K,ext}^N, \end{cases} \tag{3.2.13}$$

where $dr_{K,\sigma}^n$ is defined by (3.2.12). This flux preserves the equilibrium.

3.2.3 Consistency of the numerical flux

Lemma 3.2.1. *Let $a,\ b \in \mathbb{R},\ a, b \geq 0$. Then there exists $\eta \in [\min(a,b), \max(a,b)]$ such that*

$$dr(a,b) = r'(\eta).$$

Proof. The result is clear if $ab = 0$ or $a = b$. Let us suppose that $ab > 0$ and $a < b$ (the proof is the same if $a > b$). If we consider the change of variables $x = \log(a)$ and $y = \log(b)$, we obtain

$$dr(a,b) = \frac{h(\exp(y)) - h(\exp(x))}{y - x}$$

and using Taylor's formula, there exists $\theta \in [x, y]$ such that

$$dr(a,b) = \exp(\theta)h'(\exp(\theta)) = r'(\exp(\theta)) \text{ (using the definition of } h).$$

Finally, there exists $\eta = \exp(\theta) \in [a, b]$ such that

$$dr(a,b) = r'(\eta).$$

□

Remark 3. The flux (3.2.13) can also be written as

$$\mathcal{F}_{K,\sigma}^{n+1} = \mathrm{m}(\sigma)q_{K,\sigma}\frac{U_K^{n+1} + U_\sigma^{n+1}}{2} - \frac{\mathrm{m}(\sigma)q_{K,\sigma}}{2}\coth\left(\frac{d_\sigma q_{K,\sigma}}{2dr_{K,\sigma}^n}\right)(U_\sigma^{n+1} - U_K^{n+1}). \tag{3.2.14}$$

The first term is a centred discretization of the convective part. The second term is consistent with the diffusive part of equation (3.1.14), since $\coth(x) \underset{0}{\sim} \frac{1}{x}$.

3.3 Properties of the scheme

3.3.1 Well-posedness of the scheme

The following proposition gives the existence and uniqueness result of the solution to the scheme defined by (3.2.1)-(3.2.2)-(3.2.4)-(3.2.13) and an L^∞-estimate on this solution.

Proposition 3.3.1. *Let us assume hypotheses (H1)-(H5). Let $\mathcal{D}$ be an admissible discretization of $\Omega \times (0,T)$. Then there exists a unique solution $\{U_K^n, K \in \mathcal{T}, 0 \leq n \leq N_T\}$ to the scheme (3.2.1)-(3.2.2)-(3.2.4)-(3.2.13), with $U_K^n \geq 0$ for all $K \in \mathcal{T}$ and $0 \leq n \leq N_T$. Moreover, if we suppose that the two following assumptions are fulfilled:*

(H6) $\mathrm{div}(\mathbf{q}) = 0$,

(H7) *there exist two constants $m > 0$ and $M > 0$ such that $m \leq \overline{u}, u_0 \leq M$,*

then we have

$$0 < m \leq U_K^n \leq M, \quad \forall K \in \mathcal{T}, \quad \forall n \geq 0. \tag{3.3.1}$$

Proof. At each time step, the scheme (3.2.1)-(3.2.2)-(3.2.4)-(3.2.13) leads to a system of $\mathrm{card}(\mathcal{T})$ linear equations on $U^{n+1} = (U_K^{n+1})_{K\in\mathcal{T}}$ which can be written:

$$A^n U^{n+1} = S^n,$$

where :

- A^n is the matrix defined by

$$A_{K,K}^n = \frac{\mathrm{m}(K)}{\Delta t} + \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma dr_{K,\sigma}^n B\left(\frac{-d_\sigma q_{K,\sigma}}{dr_{K,\sigma}^n}\right) \quad \forall K \in \mathcal{T},$$

$$A_{K,L}^n = -\tau_\sigma dr_{K,\sigma}^n B\left(\frac{d_\sigma q_{K,\sigma}}{dr_{K,\sigma}^n}\right) \quad \forall L \in \mathcal{T} \text{ such that } \sigma = K|L \in \mathcal{E}_{int,K};$$

- $S^n = \left(\frac{\mathrm{m}(K)}{\Delta t} U_K^n\right)_{K\in\mathcal{T}} + Tb^n$, with

$$Tb_K^n = \begin{cases} 0 & \text{if } K \in \mathcal{T} \text{ such that } \mathrm{m}(\partial K \cap \Gamma) = 0, \\ \displaystyle\sum_{\sigma\in\mathcal{E}_{ext,K}^D} \tau_\sigma dr_{K,\sigma}^n B\left(\frac{d_\sigma q_{K,\sigma}}{dr_{K,\sigma}^n}\right) U_\sigma^{n+1} & \text{if } K \in \mathcal{T} \text{ such that } \mathrm{m}(\partial K \cap \Gamma) \neq 0. \end{cases}$$

The diagonal terms of A^n are positive and the offdiagonal terms are nonnegative (since $B(x) > 0$ for all $x \in \mathbb{R}$ and $dr_{K,\sigma}^n \geq 0$ for all $K \in \mathcal{T}$, for all $\sigma \in \mathcal{E}_K$). Moreover, since $dr_{K,\sigma}^n = dr_{L,\sigma}^n$ and $q_{K,\sigma} = -q_{L,\sigma}$ for all $\sigma = K|L \in \mathcal{E}_{int}$, we have for all $L \in \mathcal{T}$:

$$\left|A_{L,L}^n\right| - \sum_{\substack{K\in\mathcal{T} \\ K\neq L}} \left|A_{K,L}^n\right| = \frac{\mathrm{m}(L)}{\Delta t} > 0,$$

and then A^n is strictly diagonally dominant with respect to the columns. A^n is then an M-matrix so A^n is invertible, which gives existence and uniqueness of the solution of the scheme. Moreover, $(A^n)^{-1} \geq 0$ and since $U_K^0 \geq 0$ for all $K \in \mathcal{T}$ (using (H3)) and $U_\sigma^{n+1} \geq 0$ for all $n \geq 0$, for all $\sigma \in \mathcal{E}_{ext}^D$ (using (H2)), it is easy to prove by induction that $U_K^n \geq 0$ for all $K \in \mathcal{T}$, for all $n \geq 0$.
Now, we suppose that (H6) and (H7) are fulfilled. We prove that $U_K^n \leq M$ for all $K \in \mathcal{T}$, for all $n \geq 0$ by induction. Thanks to hypothesis (H7), we have clearly $U_K^0 \leq M$ for all $K \in \mathcal{T}$.
Let us suppose that $U_K^n \leq M \quad \forall K \in \mathcal{T}$. We want to prove $U_K^{n+1} \leq M \quad \forall K \in \mathcal{T}$.
Let us define $\mathbf{M} = (M, ..., M)^T \in \mathbb{R}^{\mathrm{card}(\mathcal{T})}$. Since A^n is an M-matrix, we have $(A^n)^{-1} \geq 0$ and then it suffices to prove that $A^n\left(U^{n+1} - \mathbf{M}\right) \leq 0$.
We first compute $A^n\mathbf{M}$. Using the following property of the Bernoulli function:

$$B(x) - B(-x) = -x \quad \forall x \in \mathbb{R}, \tag{3.3.2}$$

we obtain that for all $K \in \mathcal{T}$,

$$(A^n \mathbf{M})_K = M \left(\frac{\mathrm{m}(K)}{\Delta t} + \sum_{\sigma \in \mathcal{E}_{int,K}} \mathrm{m}(\sigma) q_{K,\sigma} + \sum_{\sigma \in \mathcal{E}_{ext,K}^D} \tau_\sigma dr_{K,\sigma}^n B\left(-\frac{d_\sigma q_{K,\sigma}}{dr_{K,\sigma}^n} \right) \right).$$

Then we compute $A^n \left(U^{n+1} - \mathbf{M}\right)$: for all $K \in \mathcal{T}$

$$\begin{aligned} \left(A^n \left(U^{n+1} - \mathbf{M}\right)\right)_K &= \frac{\mathrm{m}(K)}{\Delta t}(U_K^n - M) + \sum_{\sigma \in \mathcal{E}_{ext,K}^D} \tau_\sigma dr_{K,\sigma}^n B\left(\frac{d_\sigma q_{K,\sigma}}{dr_{K,\sigma}^n} \right) U_\sigma^{n+1} \\ &\quad - M \sum_{\sigma \in \mathcal{E}_{int,K}} \mathrm{m}(\sigma) q_{K,\sigma} - M \sum_{\sigma \in \mathcal{E}_{ext,K}^D} \tau_\sigma dr_{K,\sigma}^n B\left(-\frac{d_\sigma q_{K,\sigma}}{dr_{K,\sigma}^n} \right). \end{aligned}$$

By induction hypothesis, the first term is nonpositive. Moreover, using hypothesis (H7) and the property (3.3.2), we obtain

$$\begin{aligned} \left(A^n \left(U^{n+1} - \mathbf{M}\right)\right)_K &\leq -M \sum_{\sigma \in \mathcal{E}_{int,K}} \mathrm{m}(\sigma) q_{K,\sigma} - M \sum_{\sigma \in \mathcal{E}_{ext,K}^D} \mathrm{m}(\sigma) q_{K,\sigma} \\ &\leq -M \sum_{\sigma \in \mathcal{E}_K} \mathrm{m}(\sigma) q_{K,\sigma}. \end{aligned}$$

However, using hypothesis (H6) and the definition of $q_{K,\sigma}$ (3.2.3), we get

$$\sum_{\sigma \in \mathcal{E}_K} \mathrm{m}(\sigma) q_{K,\sigma} = \sum_{\sigma \in \mathcal{E}_K} \int_\sigma q \cdot \mathbf{n}_{K,\sigma} \, ds = \int_K \mathrm{div}(q) = 0,$$

and then $\left(A^n \left(U^{n+1} - \mathbf{M}\right)\right)_K \leq 0$ for all $K \in \mathcal{T}$.
So we have $A^n \left(U^{n+1} - \mathbf{M}\right) \leq 0$, therefore we deduce that $U^{n+1} - \mathbf{M} \leq 0$, hence $U_K^{n+1} \leq M \quad \forall K$ and we can show by the same way that $U_K^{n+1} \geq m \quad \forall K$. □

Remark 4. In the case of the drift-diffusion system for semiconductors, the hypothesis (H6) is not fulfilled ($\Delta V \neq 0$). Nevertheless, if we assume that

- the doping profile C is equal to 0,
- there exist two constants $m > 0$ and $M > 0$ such that $m \leq \overline{N}, N_0, \overline{P}, P_0 \leq M$,
- $M \Delta t \leq 1$,

then we have, using the same kind of proof as in [52],

$$\begin{aligned} 0 < m \leq N_K^n \leq M, \qquad \forall K \in \mathcal{T}, \quad \forall n \geq 0, \\ 0 < m \leq P_K^n \leq M, \qquad \forall K \in \mathcal{T}, \quad \forall n \geq 0. \end{aligned}$$

Definition 3.2. Let $\mathcal{D}$ be an admissible discretization of $\Omega \times (0,T)$. The approximate solution to the problem (3.1.14)-(3.1.15)-(3.1.16)-(3.1.17) associated to the discretization $\mathcal{D}$ is defined as piecewise constant function by:

$$u_\delta(x,t) = U_K^{n+1}, \quad \forall (x,t) \in K \times [t^n, t^{n+1}[, \tag{3.3.3}$$

where $\{U_K^n, K \in \mathcal{T}, 0 \leq n \leq N_T\}$ is the unique solution to the scheme (3.2.1)-(3.2.2)-(3.2.4)-(3.2.13).

3.3.2 Discrete $L^2(0,T;H^1)$ estimate on u_δ

In this section, we prove a discrete $L^2(0,T;H^1)$ estimate on u_δ in the nondegenerate case, which leads to compactness and convergence results.
For a piecewise constant function v_δ defined by $v_\delta(x,t) = v_K^{n+1}$ for $(x,t) \in K \times [t^n, t^{n+1}[$ and $v_\delta(\gamma,t) = v_\sigma^{n+1}$ for $(\gamma,t) \in \sigma \times [t^n, t^{n+1}[$, we define

$$\|v_\delta\|_{1,\mathcal{D}}^2 = \sum_{n=0}^{N_T} \Delta t \left(\sum_{\substack{\sigma \in \mathcal{E}_{int} \\ \sigma = K|L}} \tau_\sigma \left| v_L^{n+1} - v_K^{n+1} \right|^2 + \sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}_{ext,K}^D} \tau_\sigma \left| v_\sigma^{n+1} - v_K^{n+1} \right|^2 \right).$$

Proposition 3.3.2. *Let assume (H1)-(H7) are satisfied. Let u_δ be defined by the scheme (3.2.1)-(3.2.2)-(3.2.4)-(3.2.13) and (3.3.3).*
There exists $D_1 > 0$ only depending on r, $\boldsymbol{q}$, u_0, $\overline{u}$, Ω and T such that

$$\|u_\delta\|_{1,\mathcal{D}}^2 \leq D_1. \tag{3.3.4}$$

Proof. We follow the proof of Lemma 4.2 in [83]. Throughout this proof, D_i denotes constants which depend only on r, $\mathbf{q}$, u_0, $\overline{u}$, Ω and T. We set

$$\overline{U}_K^{n+1} = \frac{1}{\Delta t \mathrm{m}(K)} \int_{t^n}^{t^{n+1}} \int_K \overline{u}(x,t)\, dx\, dt,\ \forall K \in \mathcal{T},\ \ \forall n \in \mathbb{N},$$

and

$$w_K^{n+1} = U_K^{n+1} - \overline{U}_K^{n+1},\ \forall K \in \mathcal{T},\ \ \forall n \in \mathbb{N}.$$

We multiply the scheme (3.2.4) by $\Delta t w_K^{n+1}$ and we sum over n and K. We obtain $A+B=0$, where:

$$\begin{aligned} A &= \sum_{n=0}^{N_T} \sum_{K \in \mathcal{T}} \mathrm{m}(K) \left(U_K^{n+1} - U_K^n \right) w_K^{n+1}, \\ B &= \sum_{n=0}^{N_T} \Delta t \sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}_K} \mathcal{F}_{K,\sigma}^{n+1} w_K^{n+1}. \end{aligned}$$

Estimate of A. This term is treated in [83]. We get:

$$A \geq -\frac{1}{2} \|u_0 - \overline{u}(.,0)\|_{L^2(\Omega)}^2 - 2 \|\partial_t \overline{u}\|_{L^1(\Omega \times (0,T))} |M - m| = -D_2. \tag{3.3.5}$$

Estimate of B. A discrete integration by parts yields (using that $w_\sigma^{n+1} = 0$ for all $\sigma \in \mathcal{E}_{ext}^D$ and for all $n \geq 0$):

$$B = \sum_{n=0}^{N_T} \Delta t \sum_{\substack{\sigma \in \mathcal{E}_{int} \\ \sigma = K|L}} \mathcal{F}_{K,\sigma}^{n+1} \left(w_K^{n+1} - w_L^{n+1} \right) + \sum_{n=0}^{N_T} \Delta t \sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}_{ext,K}^D} \mathcal{F}_{K,\sigma}^{n+1} \left(w_K^{n+1} - w_\sigma^{n+1} \right),$$

which delivers $B = B' - \overline{B}$, with:

$$\begin{aligned} B' &= \sum_{n=0}^{N_T} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \mathcal{F}_{K,\sigma}^{n+1}\left(U_K^{n+1} - U_L^{n+1}\right) + \sum_{n=0}^{N_T} \Delta t \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{ext,K}^D} \mathcal{F}_{K,\sigma}^{n+1}\left(U_K^{n+1} - U_\sigma^{n+1}\right), \\ \overline{B} &= \sum_{n=0}^{N_T} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \mathcal{F}_{K,\sigma}^{n+1}\left(\overline{U}_K^{n+1} - \overline{U}_L^{n+1}\right) + \sum_{n=0}^{N_T} \Delta t \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{ext,K}^D} \mathcal{F}_{K,\sigma}^{n+1}\left(\overline{U}_K^{n+1} - \overline{U}_\sigma^{n+1}\right). \end{aligned}$$

Estimate of $\overline{B}$. Using the expression (3.2.14) of $\mathcal{F}_{K,\sigma}^{n+1}$, we have $\overline{B} = \overline{B}_1 + \overline{B}_2$ with

$$\begin{aligned} \overline{B}_1 &= \sum_{n=0}^{N_T} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \frac{\mathrm{m}(\sigma) q_{K,\sigma}}{2}\left(U_K^{n+1} + U_L^{n+1}\right)\left(\overline{U}_K^{n+1} - \overline{U}_L^{n+1}\right) \\ &\quad + \sum_{n=0}^{N_T} \Delta t \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{ext,K}^D} \frac{\mathrm{m}(\sigma) q_{K,\sigma}}{2}\left(U_K^{n+1} + U_\sigma^{n+1}\right)\left(\overline{U}_K^{n+1} - \overline{U}_\sigma^{n+1}\right), \\ \overline{B}_2 &= \sum_{n=0}^{N_T} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \frac{\mathrm{m}(\sigma) q_{K,\sigma}}{2}\coth\left(\frac{d_\sigma q_{K,\sigma}}{2dr_{K,\sigma}^n}\right)\left(U_K^{n+1} - U_L^{n+1}\right)\left(\overline{U}_K^{n+1} - \overline{U}_L^{n+1}\right) \\ &\quad + \sum_{n=0}^{N_T} \Delta t \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{ext,K}^D} \frac{\mathrm{m}(\sigma) q_{K,\sigma}}{2}\coth\left(\frac{d_\sigma q_{K,\sigma}}{2dr_{K,\sigma}^n}\right)\left(U_K^{n+1} - U_\sigma^{n+1}\right)\left(\overline{U}_K^{n+1} - \overline{U}_\sigma^{n+1}\right). \end{aligned}$$

The term $\overline{B}_1$ is treated like in [83], which leads to

$$|\overline{B}_1| \le M\|\mathbf{q}\|_\infty \|\overline{u}_\delta\|_{1,\mathcal{D}} \mathrm{dm}(\Omega) = D_3.$$

We apply Young's inequality for $\overline{B}_2$: for any $\alpha > 0$, we have

$$\begin{aligned} \left|\overline{B}_2\right| &\le \frac{\alpha}{2}\sum_{n=0}^{N_T} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma\left(dr_{K,\sigma}^n\right)^2 \left(\frac{d_\sigma q_{K,\sigma}}{2dr_{K,\sigma}^n}\coth\left(\frac{d_\sigma q_{K,\sigma}}{2dr_{K,\sigma}^n}\right)\right)^2 \left(U_K^{n+1} - U_L^{n+1}\right)^2 \\ &+ \frac{\alpha}{2}\sum_{n=0}^{N_T} \Delta t \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{ext,K}^D} \tau_\sigma\left(dr_{K,\sigma}^n\right)^2 \left(\frac{d_\sigma q_{K,\sigma}}{2dr_{K,\sigma}^n}\coth\left(\frac{d_\sigma q_{K,\sigma}}{2dr_{K,\sigma}^n}\right)\right)^2 \left(U_K^{n+1} - U_\sigma^{n+1}\right)^2 \\ &+ \frac{1}{2\alpha}\|\overline{u}_\delta\|_{1,\mathcal{D}}^2. \end{aligned}$$

By the hypothesis (H4), we have $\inf_{s\in[m,M]} r'(s) > 0$. Then, using Lemma 3.2.1, the L^∞ estimate on u_δ (3.3.1) and the hypothesis (H5), we have

$$\frac{d_\sigma q_{K,\sigma}}{2dr_{K,\sigma}^n} \le \frac{\|\mathbf{q}\|_\infty \mathrm{diam}(\Omega)}{\inf_{s\in[m,M]} r'(s)},\ \forall n\in\mathbb{N},\ \forall K\in\mathcal{T},\ \forall\sigma\in\mathcal{E}_K.$$

Moreover, since $x \mapsto x\coth(x)$ is continuous on $\mathbb{R}$, we obtain

$$\left(\frac{d_\sigma q_{K,\sigma}}{2dr^n_{K,\sigma}}\coth\left(\frac{d_\sigma q_{K,\sigma}}{2dr^n_{K,\sigma}}\right)\right)^2 \leq D_4,\ \forall n\in\mathbb{N},\ \forall K\in\mathcal{T},\ \forall\sigma\in\mathcal{E}_K.$$

Thus we can bound $\overline{B}$:

$$\left|\overline{B}\right| \leq D_3 + \frac{\alpha}{2}D_4\left(\sup_{s\in[m,M]} r'(s)\right)^2 \|u_\delta\|^2_{1,\mathcal{D}} + \frac{1}{2\alpha}\|\overline{u}_\delta\|_{1,\mathcal{D}}. \tag{3.3.6}$$

Estimate of B'. First, using the expression (3.2.14) of the flux and Lemma 3.2.1, we have for all $n\geq 0$, for all $K\in\mathcal{T}$ and for all $\sigma = K|L\in\mathcal{E}_{int,K}$

$$\begin{aligned}\mathcal{F}^{n+1}_{K,\sigma}\left(U^{n+1}_K - U^{n+1}_L\right) &= \frac{\mathrm{m}(\sigma)q_{K,\sigma}}{2}\left(\left(U^{n+1}_K\right)^2 - \left(U^{n+1}_L\right)^2\right)\\ &\quad +\tau_\sigma r'(\eta^n_{K,\sigma})\frac{d_\sigma q_{K,\sigma}}{2r'(\eta^n_{K,\sigma})}\coth\left(\frac{d_\sigma q_{K,\sigma}}{2r'(\eta^n_{K,\sigma})}\right)\left(U^{n+1}_K - U^{n+1}_L\right)^2.\end{aligned}$$

Then, since $x\coth(x)\geq 1$ for all $x\in\mathbb{R}$, we get:

$$\mathcal{F}^{n+1}_{K,\sigma}\left(U^{n+1}_K - U^{n+1}_L\right) \geq \frac{\mathrm{m}(\sigma)q_{K,\sigma}}{2}\left(\left(U^{n+1}_K\right)^2 - \left(U^{n+1}_L\right)^2\right) + \tau_\sigma \inf_{[m,M]} r'(s)\left(U^{n+1}_K - U^{n+1}_L\right)^2.$$

We obtain the same type of inequality for $\mathcal{F}^{n+1}_{K,\sigma}\left(U^{n+1}_K - U^{n+1}_\sigma\right)$. Thus we get

$$\begin{aligned}B' &\geq \inf_{s\in[m,M]} r'(s)\|u_\delta\|^2_{1,\mathcal{D}} + \sum_{n=0}^{N_T}\Delta t\sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}}\frac{\mathrm{m}(\sigma)q_{K,\sigma}}{2}\left(\left(U^{n+1}_K\right)^2 - \left(U^{n+1}_L\right)^2\right)\\ &\quad + \sum_{n=0}^{N_T}\Delta t\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}^D_{ext,K}}\frac{\mathrm{m}(\sigma)q_{K,\sigma}}{2}\left(\left(U^{n+1}_K\right)^2 - \left(U^{n+1}_\sigma\right)^2\right).\end{aligned}$$

Through integrating by parts and using the hypothesis (H6), we get

$$\begin{aligned}&\sum_{n=0}^{N_T}\Delta t\sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}}\frac{\mathrm{m}(\sigma)q_{K,\sigma}}{2}\left(\left(U^{n+1}_K\right)^2 - \left(U^{n+1}_L\right)^2\right)\\ +\ &\sum_{n=0}^{N_T}\Delta t\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}^D_{ext,K}}\frac{\mathrm{m}(\sigma)q_{K,\sigma}}{2}\left(\left(U^{n+1}_K\right)^2 - \left(U^{n+1}_\sigma\right)^2\right)\\ =\ &-\sum_{n=0}^{N_T}\Delta t\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}^D_{ext,K}}\frac{1}{2}\int_\sigma \mathbf{q}(x)\cdot\mathbf{n}_{K,\sigma}\,ds(x)\left(U^{n+1}_\sigma\right)^2 = -D_5,\end{aligned}$$

and then

$$B' \geq \inf_{s\in[m,M]} r'(s)\|u_\delta\|_{1,\mathcal{D}}^2 - D_5. \tag{3.3.7}$$

Conclusion. Using $A + B = 0$ and estimates (3.3.5), (3.3.6) and (3.3.7), we finally get for any $\alpha > 0$:

$$\left(\inf_{s\in[m,M]} r'(s) - \frac{\alpha}{2} D_4 \left(\sup_{s\in[m,M]} r'(s)\right)^2\right) \|u_\delta\|_{1,\mathcal{D}}^2 \leq D_2 + D_3 + D_5 + \frac{1}{2\alpha}\|\overline{u}_\delta\|_{1,\mathcal{D}}^2,$$

thus for $\alpha < \dfrac{2\inf_{s\in[m,M]} r'(s)}{D_4\left(\sup_{s\in[m,M]} r'(s)\right)^2}$, we obtain $\|u_\delta\|_{1,\mathcal{D}}^2 \leq D_1$. □

3.4 Convergence

In this section, we prove the convergence of the approximate solution u_δ to a weak solution u of the problem (3.1.14)-(3.1.15)-(3.1.16)-(3.1.17). Our first goal is to prove the strong compactness of $(u_\delta)_{\delta>0}$ in $L^2(\Omega\times]0,T[)$. It comes from the criterion of strong compactness of a sequence by using estimates (3.3.1) and (3.3.4). Then, we will prove the weak compactness in $L^2(\Omega\times]0,T[)$ of an approximate gradient. Finally, we will show the convergence of the scheme.

3.4.1 Compactness of the approximate solution

The following lemma is a classical consequence of Proposition 3.3.2 and estimates of time translation for u_δ obtained from the scheme (3.2.1)-(3.2.2)-(3.2.4)-(3.2.13). The proof is similar to those of Lemma 4.3 and Lemma 4.7 in [85].

Lemma 3.4.1 (Space and time translate estimates)**.** *We suppose (H1)-(H7). Let $\mathcal{D}$ be an admissible discretization of $\Omega\times(0,T)$. Let u_δ be defined by the scheme (3.2.1)-(3.2.2)-(3.2.4)-(3.2.13) and by (3.3.3).*
Let $\hat{u}$ be defined by $\hat{u}_\delta = u_\delta$ a.e. on $\Omega\times(0,T)$ and $\hat{u}_\delta = 0$ a.e. on $\mathbb{R}^{d+1}\setminus\Omega\times(0,T)$.
Then we get the existence of $M_2 > 0$, only depending on Ω, T, r, q, u_0, $\overline{u}$ and not on $\mathcal{D}$ such that

$$\int_0^T\int_\Omega \left(\hat{u}_\delta(x+\eta,t) - \hat{u}_\delta(x,t)\right)^2 dx\,dt \leq M_2|\eta|(|\eta| + 4\delta), \quad \forall \eta\in\mathbb{R}^d, \tag{3.4.1}$$

and

$$\int_0^T\int_\Omega \left(\hat{u}_\delta(x,t+\tau) - \hat{u}_\delta(x,t)\right)^2 dx\,dt \leq M_2|\tau|, \quad \forall \tau\in\mathbb{R}. \tag{3.4.2}$$

Now, we define an approximation $\nabla^\delta u_\delta$ of the gradient of u. Therefore, we will define a dual mesh. For $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$, we define $T_{K,\sigma}$ as follows:

- if $\sigma = K|L \in \mathcal{E}_{int,K}$, then $T_{K,\sigma}$ is the cell whose vertices are x_K, x_L and those of $\sigma = K|L$,
- if $\sigma \in \mathcal{E}_{ext,K}$, then $T_{K,\sigma}$ is the cell whose vertices are x_K and those of σ.

See [54] for an example of construction of $T_{K,\sigma}$. Then $\left((T_{K,\sigma})_{\sigma\in\mathcal{E}_K}\right)_{K\in\mathcal{T}}$ defines a partition of Ω. The approximation $\nabla^\delta u_\delta$ is a piecewise function defined in $\Omega \times (0,T)$ by:

$$\nabla^\delta u_\delta(x,t) = \begin{cases} \dfrac{\mathrm{m}(\sigma)}{\mathrm{m}(T_{K,\sigma})}\left(U_L^{n+1} - U_K^{n+1}\right)\mathbf{n}_{K,\sigma} & \text{if } (x,t) \in T_{K,\sigma} \times [t^n, t^{n+1}[,\ \sigma = K|L, \\ \dfrac{\mathrm{m}(\sigma)}{\mathrm{m}(T_{K,\sigma})}\left(U_\sigma^{n+1} - U_K^{n+1}\right)\mathbf{n}_{K,\sigma} & \text{if } (x,t) \in T_{K,\sigma} \times [t^n, t^{n+1}[,\ \sigma \in \mathcal{E}_{ext,K}. \end{cases}$$

Proposition 3.4.1. *We suppose (H1)-(H7).*
There exist subsequences of $(u_\delta)_{\delta>0}$ and $(\nabla^\delta u_\delta)_{\delta>0}$, still denoted $(u_\delta)_{\delta>0}$ and $(\nabla^\delta u_\delta)_{\delta>0}$, and a function $u \in L^\infty(0,T;H^1(\Omega))$ such that

$$\begin{array}{lll} u_\delta \to u & \text{in } L^2(\Omega\times]0,T[) \text{ strongly,} & \text{as } \delta \to 0, \\ \nabla^\delta u_\delta \rightharpoonup \nabla u & \text{in } (L^2(\Omega\times]0,T[))^d \text{ weakly,} & \text{as } \delta \to 0. \end{array}$$

Proof. Using estimates (3.4.1)-(3.4.2) and applying the Riesz-Fréchet-Kolmogorov criterion of strong compactness [31], we obtain the first part of this proposition. The result concerning $\nabla^\delta u_\delta$ is proved in [52]. □

3.4.2 Convergence of the scheme

Now it remains to prove that the function u defined in Proposition 3.4.1 satisfies Definition 3.1 of a weak solution. The main difficulty in proving this comes from the fact that the diffusive and convective terms are put together in the Scharfetter-Gummel flux.

Theorem 3.4.1. *Assume (H1)-(H7) hold. Then the function u defined in Proposition 3.4.1 satisfies the equation (3.1.14)-(3.1.15)-(3.1.16)-(3.1.17) in the sense of (3.1.18) and the boundary condition $u - \overline{u} \in L^\infty(0,T;H_0^1(\Omega))$.*

Proof. Let $\psi \in \mathcal{D}(\Omega \times [0,T[)$ be a test function and $\psi_K^n = \psi(x_K, t^n)$ for all $K \in \mathcal{T}$ and $n \geq 0$. We suppose that $\delta > 0$ is small enough such that $\mathrm{Supp}(\psi) \subset \{x \in \Omega;\ \mathrm{d}(x,\Gamma) > \delta\} \times [0, (N_T - 1)\Delta t[$. Let us define an approximate gradient of ψ by

$$\nabla^\delta \psi(x,t) = \begin{cases} \dfrac{\mathrm{m}(\sigma)}{\mathrm{m}(T_{K,\sigma})}\left(\psi_L^n - \psi_K^n\right)\mathbf{n}_{K,\sigma} & \text{if } (x,t) \in T_{K,\sigma} \times [t^n, t^{n+1}[,\ \ \sigma = K|L, \\ \dfrac{\mathrm{m}(\sigma)}{\mathrm{m}(T_{K,\sigma})}\left(\psi_\sigma^n - \psi_K^n\right)\mathbf{n}_{K,\sigma} & \text{if } (x,t) \in T_{K,\sigma} \times [t^n, t^{n+1}[,\ \ \sigma \in \mathcal{E}_{ext,K}. \end{cases}$$

We get from [84] that $(\nabla^\delta \psi)_{\delta>0}$ weakly converges to $\nabla\psi$ in $(L^2(\Omega\times(0,T)))^d$ as δ goes to zero.
Let us introduce the following notations:

$$\begin{aligned}
B_{10}(\delta) &= -\left(\int_0^T\int_\Omega u_\delta(x,t)\,\partial_t\psi(x,t)\,dx\,dt + \int_\Omega u_\delta(x,0)\,\psi(x,0)\,dx\right),\\
B_{20}(\delta) &= \int_0^T\int_\Omega r'(u_\delta(x,t-\Delta t))\,\nabla^\delta u_\delta(x,t)\cdot\nabla\psi(x,t)\,dx\,dt,\\
B_{30}(\delta) &= -\int_0^T\int_\Omega u_\delta(x,t)\,\mathbf{q}(x)\cdot\nabla^\delta\psi(x,t)\,dx\,dt,
\end{aligned}$$

and

$$\varepsilon(\delta) = -B_{10}(\delta) - B_{20}(\delta) - B_{30}(\delta).$$

Multiplying the scheme (3.2.4) by $\Delta t\psi_K^n$ and summing through K and n, we obtain

$$B_1(\delta) + B_2(\delta) + B_3(\delta) = 0,$$

where

$$\begin{aligned}
B_1(\delta) &= \sum_{n=0}^{N_T}\sum_{K\in\mathcal{T}} \mathrm{m}(K)\left(U_K^{n+1}-U_K^n\right)\psi_K^n,\\
B_2(\delta) &= -\sum_{n=0}^{N_T}\Delta t\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}\frac{\mathrm{m}(\sigma)q_{K,\sigma}}{2}\coth\left(\frac{d_\sigma q_{K,\sigma}}{2dr_{K,\sigma}^n}\right)\left(U_\sigma^{n+1}-U_K^{n+1}\right)\psi_K^n,\\
B_3(\delta) &= \sum_{n=0}^{N_T}\Delta t\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}\mathrm{m}(\sigma)q_{K,\sigma}\frac{U_K^{n+1}+U_\sigma^{n+1}}{2}\psi_K^n.
\end{aligned}$$

From the strong convergence of the sequence $(u_\delta)_{\delta>0}$ to u in $L^2(\Omega\times]0,T[)$, it is clear using the time translate estimate (3.4.2) that there exists a subsequence of $(u_\delta)_{\delta>0}$, still denoted by $(u_\delta)_{\delta>0}$, such that

$$u_\delta(\,\cdot\,,\,\cdot\,-\Delta t)\longrightarrow u \text{ in } L^2(\Omega\times]0,T[) \text{ strongly as } \delta\to 0,$$

where $u\in L^\infty(0,T;H^1(\Omega))$ is defined in Proposition 3.4.1. Moreover, thanks to hypothesis (H4), we have $r'\in\mathcal{C}^1(\mathbb{R})$, and using the L^∞-estimate (3.3.1) we obtain that

$$r'(u_\delta(\,\cdot\,,\,\cdot\,-\Delta t))\longrightarrow r'(u) \text{ in } L^2(\Omega\times]0,T[) \text{ strongly as } \delta\to 0.$$

Finally using this strong convergence and the weak convergence of the sequences $(\nabla^\delta u_\delta)_{\delta>0}$ to ∇u and $(\nabla^\delta\psi)_{\delta>0}$ to $\nabla\psi$ in $(L^2(\Omega\times]0,T[))^d$, it is easy to see that

$$\begin{aligned}
\varepsilon(\delta) \longrightarrow\ & \int_0^T\int_\Omega \left(u(x,t)\,\partial_t\psi - r'(u(x,t))\,\nabla u(x,t)\cdot\nabla\psi + u(x,t)\,\mathbf{q}(x)\cdot\nabla\psi\right)\,dx\,dt\\
&+\int_\Omega u(x,0)\,\psi(x,0)\,dx, \text{ as } \delta\to 0.
\end{aligned}$$

Therefore, it suffices to prove that $\varepsilon(\delta) \longrightarrow 0$ as $\delta \to 0$ and to this end we are going to prove that $\varepsilon(\delta) + B_1(\delta) + B_2(\delta) + B_3(\delta) \longrightarrow 0$ as $\delta \to 0$.

Estimate of $B_1(\delta) - B_{10}(\delta)$. This term is discussed for example in [52, Theorem 5.2] and it is proved that:

$$|B_1(\delta) - B_{10}(\delta)| \leq \left[(T+1)\mathrm{m}(\Omega)M\|\psi\|_{\mathcal{C}^2(\Omega\times(0,T))}\right]\delta \longrightarrow 0 \text{ as } \delta \to 0.$$

Estimate of $B_2(\delta) - B_{20}(\delta)$. Using a discrete integration by parts, we write

$$B_2(\delta) = \sum_{n=0}^{N_T} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \frac{\mathrm{m}(\sigma)q_{K,\sigma}}{2} \coth\left(\frac{d_\sigma q_{K,\sigma}}{2dr^n_{K,\sigma}}\right)\left(U_L^{n+1} - U_K^{n+1}\right)(\psi_L^n - \psi_K^n).$$

Then we rewrite $B_2(\delta) = B_{21}(\delta) + B_{22}(\delta) + B_{23}(\delta)$, with

$$B_{21}(\delta) = \sum_{n=0}^{N_T} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma r'(U_K^n)\left(U_L^{n+1} - U_K^{n+1}\right)(\psi_L^n - \psi_K^n),$$

$$B_{22}(\delta) = \sum_{n=0}^{N_T} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma \left(\frac{d_\sigma q_{K,\sigma}}{2dr^n_{K,\sigma}} \coth\left(\frac{d_\sigma q_{K,\sigma}}{2dr^n_{K,\sigma}}\right) - 1\right) dr^n_{K,\sigma}\left(U_L^{n+1} - U_K^{n+1}\right)(\psi_L^n - \psi_K^n),$$

$$B_{23}(\delta) = \sum_{n=0}^{N_T} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma \left(dr^n_{K,\sigma} - r'(U_K^n)\right)\left(U_L^{n+1} - U_K^{n+1}\right)(\psi_L^n - \psi_K^n).$$

Using the definition of $\tilde{u}_\delta$ and $\nabla^\delta u_\delta$, we rewrite $B_{20}(\delta)$ as $B_{210}(\delta) + B_{220}(\delta)$ with:

$$B_{210}(\delta) = \sum_{n=0}^{N_T} \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} r'(U_K^n)\frac{\mathrm{m}(\sigma)}{\mathrm{m}(T_{K,\sigma})}\left(U_L^{n+1} - U_K^{n+1}\right)\int_{t^n}^{t^{n+1}}\int_{T_{K,\sigma}} \nabla\psi(x,t)\cdot\mathbf{n}_{K,\sigma}\,dx\,dt,$$

$$B_{220}(\delta) = \sum_{n=0}^{N_T} \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \left(r'(U_L^n) - r'(U_K^n)\right)\frac{\mathrm{m}(\sigma)}{\mathrm{m}(T_{K,\sigma})}\left(U_L^{n+1} - U_K^{n+1}\right)$$
$$\times \int_{t^n}^{t^{n+1}}\int_{T_{K,\sigma}\cap L} \nabla\psi(x,t)\cdot\mathbf{n}_{K,\sigma}\,dx\,dt.$$

Now we prove that $B_{21}(\delta) - B_{210}(\delta) \to 0$ as $\delta \to 0$ and $B_{22}(\delta)$, $B_{23}(\delta)$, $B_{220}(\delta) \to 0$ as $\delta \to 0$.

Estimate of $B_{21}(\delta) - B_{210}(\delta)$. We have

$$B_{21}(\delta) - B_{210}(\delta) = \sum_{n=0}^{N_T} \sum_{\sigma \in \mathcal{E}_{int}} \mathrm{m}(\sigma) r'(U_K^n) \times \left[\int_{t^n}^{t^{n+1}} \left(\frac{\psi_L^n - \psi_K^n}{d_\sigma} - \frac{1}{\mathrm{m}(T_{K,\sigma})} \int_{T_{K,\sigma}} \nabla \psi(x,t) \cdot \mathbf{n}_{K,\sigma} \, dx \right) dt \right].$$

Since the straight line $\overline{x_K x_L}$ is orthogonal to the edge $K|L$, we have $x_L - x_K = d_\sigma \mathbf{n}_{K,\sigma}$ and then from the regularity of ψ,

$$\begin{aligned} \frac{\psi_L^n - \psi_K^n}{d_\sigma} &= \nabla \psi(x_K, t^n) \cdot \mathbf{n}_{K,\sigma} + O(\Delta x) \\ &= \nabla \psi(x,t) \cdot \mathbf{n}_{K,\sigma} + O(\delta), \ \ \forall (x,t) \in T_{K,\sigma} \times \left(t^n, t^{n+1} \right). \end{aligned}$$

Then by taking the mean value over $T_{K,\sigma}$, there exists $D_6 > 0$ depending only on ψ such that

$$\left| \int_{t^n}^{t^{n+1}} \left(\frac{\psi_L^n - \psi_K^n}{d_\sigma} - \frac{1}{\mathrm{m}(T_{K,\sigma})} \int_{T_{K,\sigma}} \nabla \psi \cdot \mathbf{n}_{K,\sigma} \, dx \right) dt \right| \leq D_6 \delta \Delta t,$$

and then

$$|B_{21}(\delta) - B_{210}(\delta)| \leq \delta D_6 \sup_{s \in [m,M]} r'(s) \sum_{n=0}^{N_T} \Delta t \sum_{\sigma \in \mathcal{E}_{int}} \mathrm{m}(\sigma) \left| U_L^{n+1} - U_K^{n+1} \right|.$$

Since the straight line $\overline{x_K x_L}$ is orthogonal to the edge $\sigma = K|L$ for all $\sigma \in \mathcal{E}_{int,K}$ and the mesh is regular, there is a constant $D_7 > 0$ depending only on the dimension of the domain and the geometry of $\mathcal{T}$ such that $\mathrm{m}(\sigma) d_\sigma \leq D_7 \mathrm{m}(T_{K,\sigma})$ for all $K \in \mathcal{T}$, all $\sigma \in \mathcal{E}_{ext,K}$ and then using the Cauchy-Schwarz inequality and the $L^2(0,T;H^1)$ estimate (3.3.4), we obtain

$$|B_{21}(\delta) - B_{210}(\delta)| \leq \delta D_6 \sup_{s \in [m,M]} r'(s) \sqrt{D_1 T D_7 \mathrm{m}(\Omega)} \longrightarrow 0 \text{ as } \delta \to 0.$$

Estimate of $B_{22}(\delta)$. Since $x \mapsto x \coth(x)$ is a 1-Lipschitz continuous function and is equal to 1 in 0, we have

$$\begin{aligned} |B_{22}(\delta)| &\leq \sum_{n=0}^{N_T} \Delta t \sum_{\sigma \in \mathcal{E}_{int}} \frac{\mathrm{m}(\sigma)}{2} |q_{K,\sigma}| \left| U_L^{n+1} - U_K^{n+1} \right| |\psi_L^n - \psi_K^n| \\ &\leq 2\delta \|\mathbf{q}\|_\infty \sum_{n=0}^{N_T} \Delta t \sum_{\sigma \in \mathcal{E}_{int}} \tau_\sigma \left| U_L^{n+1} - U_K^{n+1} \right| |\psi_L^n - \psi_K^n|, \text{ since } d_\sigma \leq 2\delta. \end{aligned}$$

Then using the Cauchy-Schwarz inequality, the regularity of ψ and the $L^2(0,T;H^1)$ estimate (3.3.4), there exists $D_8 > 0$ only depending on T and Ω such that:

$$|B_{22}(\delta)| \leq \delta \|\mathbf{q}\|_\infty D_8 \|\psi\|_{\mathcal{C}^1} \sqrt{D_1} \longrightarrow 0 \text{ as } \delta \to 0.$$

Estimate of $B_{23}(\delta)$. Using Lemma 3.2.1 and hypothesis (H4), we have

$$\left|dr_{K,\sigma}^n - r'(U_K^n)\right| \leq \sup_{s\in[m,M]} |r''(s)|\,|U_L^n - U_K^n|\,,\ \ \forall \sigma \in \mathcal{E}_{int},\ \ \sigma = K|L.$$

Using the regularity of ψ and the Cauchy-Schwarz inequality, we obtain

$$|B_{23}(\delta)| \leq \delta \sup_{s\in[m,M]} |r''(s)|\|\psi\|_{\mathcal{C}^1} \sum_{n=0}^{N_T} \Delta t \sum_{\sigma\in\mathcal{E}_{int}} \tau_\sigma\,|U_L^n - U_K^n|\left|U_L^{n+1} - U_K^{n+1}\right|,$$

and then using the $L^2(0,T;H^1)$ estimate (3.3.4), we get

$$|B_{23}(\delta)| \leq \delta \sup_{s\in[m,M]} |r''(s)|\|\psi\|_{\mathcal{C}^1} D_1 \longrightarrow 0 \text{ as } \delta \to 0.$$

Estimate of $B_{220}(\delta)$. We obtain the same type of estimate as for $B_{23}(\delta)$:

$$|B_{220}(\delta)| \leq 2\delta \sup_{s\in[m,M]} |r''(s)|\|\psi\|_{\mathcal{C}^1} D_1 \longrightarrow 0 \text{ as } \delta \to 0.$$

Estimate of $B_3(\delta) - B_{30}(\delta)$. Using a discrete integration by parts, we obtain

$$B_3(\delta) = -\sum_{n=0}^{N_T} \Delta t \sum_{\sigma\in\mathcal{E}_{int}} \mathrm{m}(\sigma) q_{K,\sigma} \frac{U_K^{n+1} + U_L^{n+1}}{2}\,(\psi_L^n - \psi_K^n)\,,$$

and then we rewrite $B_3(\delta)$ as $B_{31}(\delta) + B_{32}(\delta)$, with

$$\begin{aligned} B_{31}(\delta) &= -\sum_{n=0}^{N_T} \Delta t \sum_{\sigma\in\mathcal{E}_{int}} \mathrm{m}(\sigma) q_{K,\sigma} \frac{U_L^{n+1} - U_K^{n+1}}{2}\,(\psi_L^n - \psi_K^n)\,, \\ B_{32}(\delta) &= -\sum_{n=0}^{N_T} \Delta t \sum_{\sigma\in\mathcal{E}_{int}} \mathrm{m}(\sigma) q_{K,\sigma} U_K^{n+1}\,(\psi_L^n - \psi_K^n)\,. \end{aligned}$$

Using the definition of $\nabla^\delta \psi$, we get

$$B_{30}(\delta) = -\sum_{n=0}^{N_T} \sum_{\sigma\in\mathcal{E}_{int}} \int_{t^n}^{t^{n+1}} \int_{T_{K,\sigma}} u_\delta(x,t) \frac{\mathrm{m}(\sigma)}{\mathrm{m}(T_{K,\sigma})}\,(\psi_L^n - \psi_K^n)\,\mathbf{q}(x)\cdot\mathbf{n}_{K,\sigma}\,dx\,dt,$$

which gives, using the definition of u_δ, $B_{30}(\delta) = B_{310}(\delta) + B_{320}(\delta)$, where

$$B_{310}(\delta) = -\sum_{n=0}^{N_T} \Delta t \sum_{\sigma\in\mathcal{E}_{int}} \mathrm{m}(\sigma) \left(U_L^{n+1} - U_K^{n+1}\right)(\psi_L^n - \psi_K^n)\,\frac{1}{\mathrm{m}(T_{K,\sigma})} \int_{T_{K,\sigma}\cap L} \mathbf{q}(x)\cdot\mathbf{n}_{K,\sigma}\,dx,$$

$$B_{320}(\delta) = -\sum_{n=0}^{N_T} \sum_{\sigma\in\mathcal{E}_{int}} \mathrm{m}(\sigma) U_K^{n+1}\,(\psi_L^n - \psi_K^n)\,\frac{1}{\mathrm{m}(T_{K,\sigma})} \int_{T_{K,\sigma}} \mathbf{q}(x)\cdot\mathbf{n}_{K,\sigma}\,dx.$$

Now we prove that $B_{32}(\delta) - B_{320}(\delta) \to 0$ as $\delta \to 0$ and $B_{31}(\delta)$, $B_{310}(\delta) \to 0$ as $\delta \to 0$. Using the regularity of $\mathbf{q}$, there exists $D_9 > 0$ which does not depend on δ such that

$$\left| \frac{1}{\mathrm{m}(\sigma)} \int_\sigma \mathbf{q}(x) \cdot \mathbf{n}_{K,\sigma}\, ds(x) - \frac{1}{\mathrm{m}(T_{K,\sigma})} \int_{T_{K,\sigma}} \mathbf{q}(x) \cdot \mathbf{n}_{K,\sigma}\, dx \right| \leq D_9 \delta.$$

Then we can estimate $B_{32}(\delta) - B_{320}(\delta)$:

$$\begin{aligned} |B_{32}(\delta) - B_{320}(\delta)| &\leq \delta D_9 M \sum_{n=0}^{N_T} \Delta t \sum_{\sigma \in \mathcal{E}_{int}} \mathrm{m}(\sigma)\, |\psi_L^n - \psi_K^n| \\ &\leq \delta D_8 D_9 M \|\psi\|_{\mathcal{C}^1} \sqrt{D_7 \mathrm{m}(\Omega)} \longrightarrow 0 \text{ as } \delta \to 0. \end{aligned}$$

Moreover, we have

$$\begin{aligned} |B_{31}(\delta)| &\leq \delta \|\mathbf{q}\|_\infty \sum_{n=0}^{N_T} \Delta t \sum_{\sigma \in \mathcal{E}_{int}} \tau_\sigma \left| U_L^{n+1} - U_K^{n+1} \right| |\psi_L^n - \psi_K^n| \\ &\leq \delta \|\mathbf{q}\|_\infty \|\psi\|_{\mathcal{C}^1} D_8 \sqrt{D_1} \longrightarrow 0 \text{ as } \delta \to 0. \end{aligned}$$

We obtain in the same way that $B_{310}(\delta) \longrightarrow 0$ as $\delta \to 0$.
Hence u satisfies

$$\begin{aligned} &\int_0^T \int_\Omega \left(u(x,t)\, \partial_t \psi(x,t) + r'(u(x,t))\, \nabla u(x,t) \cdot \nabla \psi(x,t) + u(x,t)\, \mathbf{q}(x) \cdot \nabla \psi(x,t) \right) dx\, dt \\ &+ \int_\Omega u(x,0)\, \psi(x,0)\, dx = 0, \end{aligned}$$

and then

$$\begin{aligned} &\int_0^T \int_\Omega \left(u(x,t)\, \partial_t \psi(x,t) + \nabla (r(u(x,t))) \cdot \nabla \psi(x,t) + u(x,t)\, \mathbf{q}(x) \cdot \nabla \psi(x,t) \right) dx\, dt \\ &+ \int_\Omega u(x,0)\, \psi(x,0)\, dx = 0. \end{aligned}$$

It remains to show that $u - \overline{u} \in L^\infty(0,T; H_0^1(\Omega))$. This proof is based on the $L^2(0,T;H^1)$ estimate (3.3.4) and is similar to the one of Theorem 5.1 in [52].

□

3.5 Numerical simulations

3.5.1 Order of convergence

We consider the following one dimensional test case, picked in the paper of R. Eymard, J. Fuhrmann and K. Gärtner [83]. We look at the case where, in (3.1.14) we take $\Omega = (0,1)$,

$T = 0.004$, $r : s \mapsto s^2$, $q = 100$, in (3.1.15) we take $u_0 = 0$ and in (3.1.16) we take, for $v = 200$,

$$\begin{aligned} \overline{u}(0,t) &= (v-q)vt/2 \\ \overline{u}(1,t) &= \begin{cases} 0 & \text{for } t < 1/v, \\ (v-q)(vt-1)/2 & \text{otherwise.} \end{cases} \end{aligned}$$

The unique weak solution of this problem is then given by

$$u(x,t) = \begin{cases} (v-q)(vt-x)/2 & \text{if } x < vt, \\ 0 & \text{if } x \geq vt. \end{cases}$$

The time step is taken equal to $\Delta t = 10^{-8}$ to study the order of convergence with respect to the spatial step size Δx. In Tables 3.1 and 3.2, we compare the order of convergence in L^∞ and L^2 norms of the scheme (3.2.1)-(3.2.2)-(3.2.4) defined on one hand with the classical upwind flux (3.2.5) and on the other hand with the Scharfetter-Gummel extended flux (3.2.13). We obtain the same order of convergence as in [83]. Moreover, it appears that even if we are in a degenerate case, the Scharfetter-Gummel extended scheme is more accurate than the classical upwind scheme.

j	$\Delta x(j)$	$\|u-u_\delta\|_{L^\infty}$ Upwind	Order	$\|u-u_\delta\|_{L^\infty}$ SG extended	Order
0	$2.5.10^{-2}$	1.110		$2.137.10^{-1}$	
1	$1.25.10^{-2}$	$7.237.10^{-1}$	0.62	$1.107.10^{-1}$	0.95
2	$6.3.10^{-3}$	$4.485.10^{-1}$	0.69	$5.631.10^{-2}$	0.98
3	$3.1.10^{-3}$	$2.685.10^{-1}$	0.74	$2.84.10^{-2}$	0.99
4	$1.6.10^{-3}$	$1.568.10^{-1}$	0.78	$1.426.10^{-2}$	1
5	8.10^{-4}	9.10^{-2}	0.80	$7.15.10^{-3}$	1

Table 3.1: Experimental order of convergence in L^∞ norm for spatial step sizes $\Delta x(j) = \dfrac{0.1}{2^{j+2}}$ of the classical upwind scheme and of the Scharfetter-Gummel extended scheme.

3.5.2 Large time behavior

The drift-diffusion system for semiconductors

We may define the finite volume approximation of the drift-diffusion system (3.1.1). Initial and boundary conditions are approximated by (3.2.1) and (3.2.2). The doping profile is approximated by $(C_K)_{K\in\mathcal{T}}$ by taking the mean value of C on each volume K.

j	$\Delta x(j)$	$\|u-u_\delta\|_{L^2}$ Upwind	Order	$\|u-u_\delta\|_{L^2}$ SG extended	Order
0	$2.5.10^{-2}$	$3.336.10^{-1}$		$4.806.10^{-2}$	
1	$1.25.10^{-2}$	$1.852.10^{-1}$	0.85	$1.642.10^{-2}$	1.55
2	$6.3.10^{-3}$	$9.911.10^{-2}$	0.9	$5.695.10^{-3}$	1.53
3	$3.1.10^{-3}$	$5.182.10^{-2}$	0.94	2.10^{-3}	1.51
4	$1.6.10^{-3}$	$2.669.10^{-2}$	0.96	$7.142.10^{-4}$	1.49
5	8.10^{-4}	$1.361.10^{-2}$	0.97	$2.695.10^{-4}$	1.41

Table 3.2: Experimental order of convergence in L^2 norm for spatial step sizes $\Delta x(j) = \frac{0.1}{2^{j+2}}$ of the classical upwind scheme and of the Scharfetter-Gummel extended scheme.

The scheme for the system (3.1.1) is given by:

$$\begin{cases} \mathrm{m}(K)\dfrac{N_K^{n+1}-N_K^n}{\Delta t} + \displaystyle\sum_{\sigma\in\mathcal{E}_K} \mathcal{F}_{K,\sigma}^{n+1} = 0, & \forall K\in\mathcal{T}, \forall n\geq 0, \\ \mathrm{m}(K)\dfrac{P_K^{n+1}-P_K^n}{\Delta t} + \displaystyle\sum_{\sigma\in\mathcal{E}_K} \mathcal{G}_{K,\sigma}^{n+1} = 0, & \forall K\in\mathcal{T}, \forall n\geq 0, \\ \displaystyle\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\, DV_{K,\sigma}^n = \mathrm{m}(K)\left(N_K^n - P_K^n - C_K\right), & \forall K\in\mathcal{T}, \forall n\geq 0, \end{cases}$$

where

$$\mathcal{F}_{K,\sigma}^{n+1} = \tau_\sigma dr\left(N_K^n, N_\sigma^n\right)\left(B\left(\frac{-DV_{K,\sigma}^n}{dr\left(N_K^n,N_\sigma^n\right)}\right)N_K^{n+1} - B\left(\frac{DV_{K,\sigma}^n}{dr(N_K^n,N_\sigma^n)}\right)N_\sigma^{n+1}\right), \forall\sigma\in\mathcal{E}_K,$$

and

$$\mathcal{G}_{K,\sigma}^{n+1} = \tau_\sigma dr(P_K^n,P_\sigma^n)\left(B\left(\frac{DV_{K,\sigma}^n}{dr(P_K^n,P_\sigma^n)}\right)P_K^{n+1} - B\left(\frac{-DV_{K,\sigma}^n}{dr(P_K^n,P_\sigma^n)}\right)P_\sigma^{n+1}\right), \forall\sigma\in\mathcal{E}_K.$$

We compute an approximation $(N_K^{eq}, P_K^{eq}, V_K^{eq})_{K\in\mathcal{T}}$ of the thermal equilibrium defined by (3.1.3)-(3.1.4) with the finite volume scheme proposed by C. Chainais-Hillairet and F. Filbet in [51].
Then we introduce the discrete version of the deviation of the total energy from the thermal equilibrium (3.1.6): for $n\geq 0$,

$$\begin{aligned} \mathcal{E}^n \;=\; & \sum_{K\in\mathcal{T}} \mathrm{m}(K)\left(H(N_K^n) - H(N_K^{eq}) - h(N_K^{eq})\left(N_K^n - N_K^{eq}\right)\right) \\ & + \sum_{K\in\mathcal{T}} \mathrm{m}(K)\left(H(P_K^n) - H(P_K^{eq}) - h(P_K^{eq})(P_K^n - P_K^{eq})\right) \\ & + \frac{1}{2}\sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma\left|DV_{K,\sigma}^n - DV_{K,\sigma}^{eq}\right|^2 + \frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_{ext,K}^D} \tau_\sigma\left|DV_{K,\sigma}^n - DV_{K,\sigma}^{eq}\right|^2, \end{aligned}$$

and the discrete version of the energy dissipation (3.1.7): for $n \geq 0$,

$$\begin{aligned}\mathcal{I}^n &= \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma \min\left(N_K^{n+1}, N_L^{n+1}\right)\left[D\left(h\left(N^{n+1}\right)-V^n\right)_{K,\sigma}\right]^2 \\ &+ \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_{ext,K}} \tau_\sigma \min\left(N_K^{n+1}, N_\sigma^{n+1}\right)\left[D\left(h\left(N^{n+1}\right)-V^n\right)_{K,\sigma}\right]^2 \\ &+ \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma \min\left(P_K^{n+1}, P_L^{n+1}\right)\left[D\left(h\left(P^{n+1}\right)+V^n\right)_{K,\sigma}\right]^2 \\ &+ \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_{ext,K}} \tau_\sigma \min\left(P_K^{n+1}, P_\sigma^{n+1}\right)\left[D\left(h\left(P^{n+1}\right)+V^n\right)_{K,\sigma}\right]^2.\end{aligned}$$

We present a test case for a geometry corresponding to a PN-junction in 2D picked in the paper of C. Chainais-Hillairet and F. Filbet [51]. The doping profile is piecewise constant, equal to $+1$ in the N-region and -1 in the P-region.
The Dirichlet boundary conditions are

$$\begin{aligned}&\overline{N}=0.1,\ \ \overline{P}=0.9,\ \ \overline{V}=\frac{h(\overline{N})-h(\overline{P})}{2} \qquad \text{on } \{y=1,\ \ 0\leq x\leq 0.25\},\\ &\overline{N}=0.9,\ \ \overline{P}=0.1,\ \ \overline{V}=\frac{h(\overline{N})-h(\overline{P})}{2} \qquad \text{on } \{y=0\}.\end{aligned}$$

Elsewhere, we put homogeneous Neumann boundary conditions.
The pressure is nonlinear: $r(s)=s^\gamma$ with $\gamma=5/3$, which corresponds to the isentropic model.
We compute the numerical approximation of the thermal equilibrium and of the transient drift-diffusion system on a mesh made of 896 triangles, with time step $\Delta t=0.01$.
We then compare the large time behavior of approximate solutions obtained with the three following fluxes:

- the upwind flux defined by (3.2.5) (**Upwind**),
- the Scharfetter-Gummel extended flux (3.2.13) with the first choice (3.2.9) of $dr_{K,\sigma}$, close to that of Jüngel and Pietra (**SG-JP**),
- the Scharfetter-Gummel extended flux (3.2.13) with the new definition (3.2.11) of $dr_{K,\sigma}$ (**SG-ext**).

In Figure 3.3 we compare the discrete relative energy $\mathcal{E}^n$ and its dissipation $\mathcal{I}^n$ obtained with the **Upwind** flux, the **SG-JP** flux and the **SG-ext** flux. With the third scheme, we observe that $\mathcal{E}^n$ and $\mathcal{I}^n$ converge to zero when time goes to infinity, without a saturation phenomenon. This scheme is the only one of the three which preserves thermal equilibrium, so it appears that this property is crucial to have a good asymptotic behavior.
In Figure 3.4 we compare the relative energy $\mathcal{E}^n$ and its dissipation $\mathcal{I}^n$ obtained with the **SG-ext** flux for three different time steps $\Delta t = 5.10^{-3}$, 10^{-3}, 10^{-4}. It appears that the decay rate does not depend on the time step.

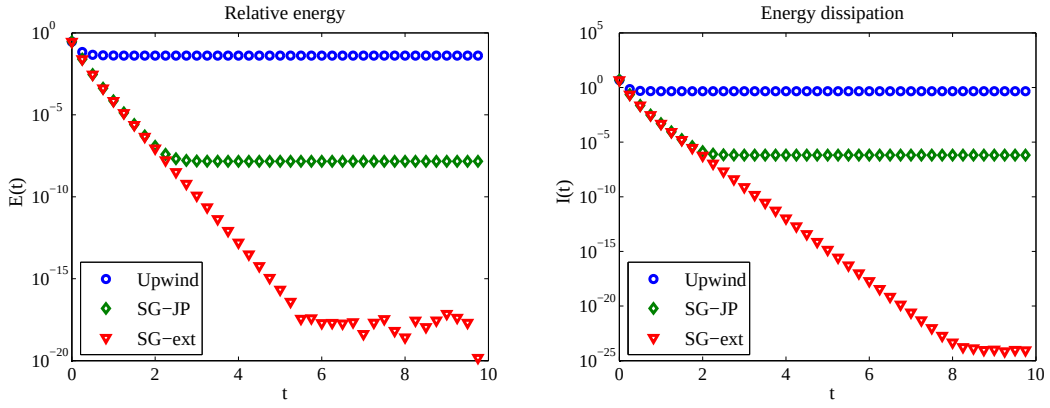


Figure 3.3: Evolution of the relative energy $\mathcal{E}^n$ and its dissipation $\mathcal{I}^n$ in log-scale for different schemes.

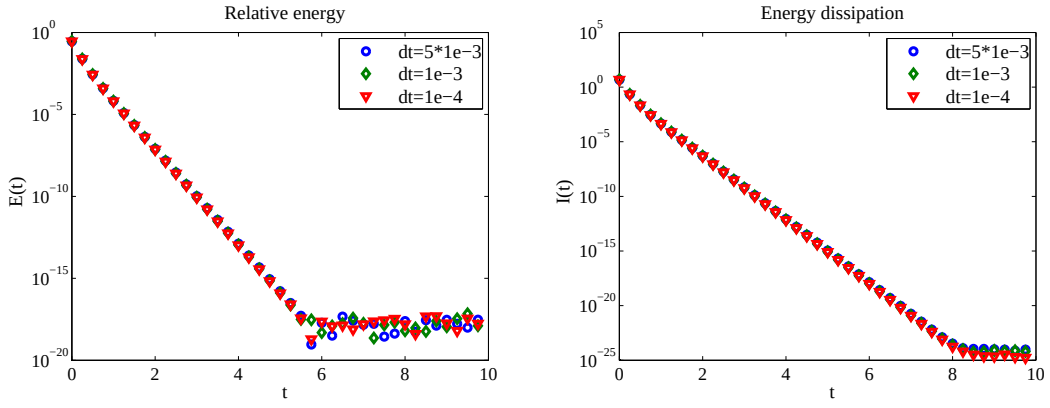


Figure 3.4: The relative energy $\mathcal{E}^n$ and its dissipation $\mathcal{I}^n$ in log-scale for different time steps.

The porous media equation

We recall that the unique stationary solution u^{eq} of the porous media equation (3.1.10) is given by the Barenblatt-Pattle type formula (3.1.11), where C_1 is such that u^{eq} as the same mass as the initial data u_0. We define an approximation $(U_K^{eq})_{K\in\mathcal{T}}$ of u^{eq} by

$$U_K^{eq} = \left(\tilde{C}_1 - \frac{\gamma-1}{2\gamma}\,|x_K|^2\right)_+^{1/(\gamma-1)}, \;\; K \in \mathcal{T},$$

where $\tilde{C}_1$ is such that the discrete mass of $(U_K^{eq})_{K\in\mathcal{T}}$ is equal to that of $\left(U_K^0\right)_{K\in\mathcal{T}}$, namely $\sum_{K\in\mathcal{T}} \mathrm{m}(K)U_K^{eq} = \sum_{K\in\mathcal{T}} \mathrm{m}(K)U_K^0$. We use a fixed point algorithm to compute this constant $\tilde{C}_1$.

We introduce the discrete version of the relative entropy (3.1.12)

$$\mathcal{E}^n = \sum_{K\in\mathcal{T}} \mathrm{m}(K)\left(H(U_K^n) - H(U_K^{eq}) + \frac{|x_K|^2}{2}(U_K^n - U_K^{eq})\right),$$

and the discrete version of the entropy dissipation (3.1.13)

$$\begin{aligned}\mathcal{I}^n \;=\; & \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma \min\left(U_K^n, U_L^n\right)\left|D\left(h(U^n) + \frac{|x|^2}{2}\right)_{K,\sigma}\right|^2 \\ & + \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_{ext,K}} \tau_\sigma \min\left(U_K^n, U_\sigma^n\right)\left|D\left(h(U^n) + \frac{|x|^2}{2}\right)_{K,\sigma}\right|^2.\end{aligned}$$

We consider the following two dimensional test case: $r(s) = s^3$, with initial condition

$$u_0(x,y) = \begin{cases} \exp\left(-\frac{1}{6-(x-2)^2-(y+2)^2}\right) & \text{if} \quad (x-2)^2+(y+2)^2 < 6, \\ \exp\left(-\frac{1}{6-(x+2)^2-(y-2)^2}\right) & \text{if} \quad (x+2)^2+(y-2)^2 < 6, \\ 0 & \text{otherwise,} \end{cases}$$

and periodic boundary conditions.

Then we compute the approximate solution on $\Omega\times(0,10)$ with $\Omega = (-10,10)\times(-10,10)$. We consider a uniform cartesian grid with 100×100 points and the time step is fixed to $\Delta t = 5.10^{-4}$.

In Figure 3.5, we plot the evolution of the numerical solution u computed with the **SG-ext** flux at three different times $t = 0$, $t = 0.4$ and $t = 4$ and the approximation of the Barenblatt-Pattle solution. In Figure 3.6 we compare the relative entropy $\mathcal{E}^n$ and its dissipation $\mathcal{I}^n$ computed with the scheme (3.2.4) and different fluxes: the **Upwind** flux, the **SG-JP** flux and the **SG-ext** flux. We made the same findings as in the case of the

drift-diffusion system for semiconductors: the third scheme is the only one of the three for which there is no saturation phenomenon, which confirms the importance of preserving the equilibrium to obtain a consistent asymptotic behavior of the approximate solution. Moreover it appears that the entropy decays exponentially fast, which has been proved in [48].
In Figure 3.7, we represent the discrete L^1 norm of $U - U^{eq}$ (obtained with the **SG-ext** flux) in log scale. According to the paper of J. A. Carrillo and G. Toscani, there exists a constant $C > 0$ such that, in this case,

$$\|u(t,x) - u^{eq}(x)\|_{L^1(\mathbb{R})} \leq C \exp\left(-\frac{3}{5}t\right), \quad t \geq 0.$$

We observe that the experimental decay of u towards the steady state u^{eq} is exponential, at a rate better than 3/5.

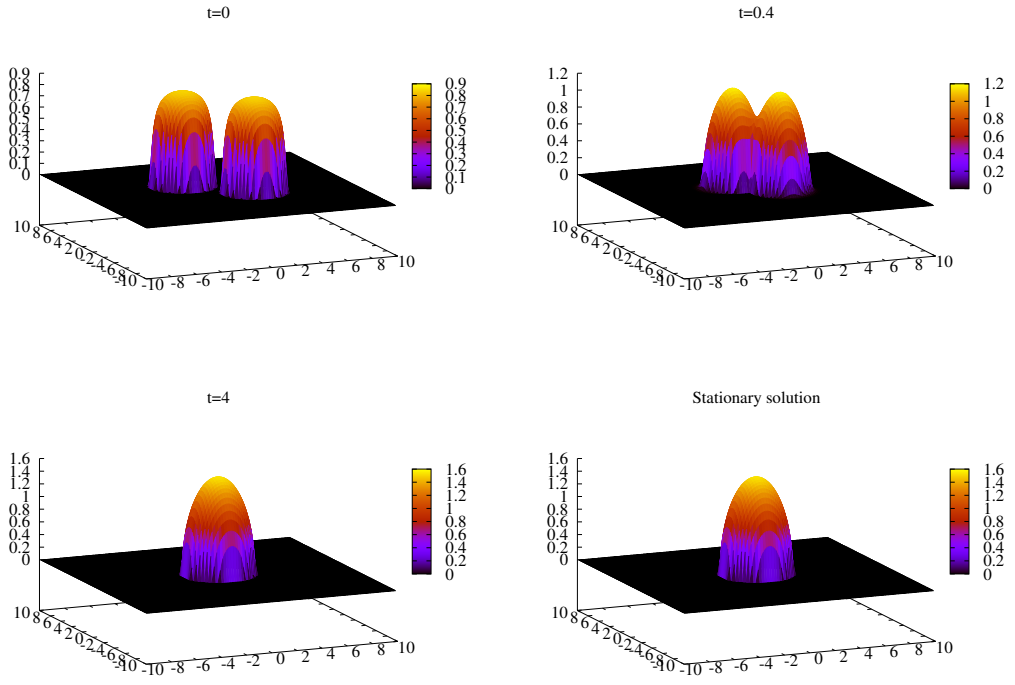


Figure 3.5: Evolution of the density of the gas u and stationary solution u^{eq}.

3.6 Conclusion

In this chapter, we presented how to build a new finite volume scheme for nonlinear convection-diffusion equations. To this end, we have to adapt the Scharfetter-Gummel

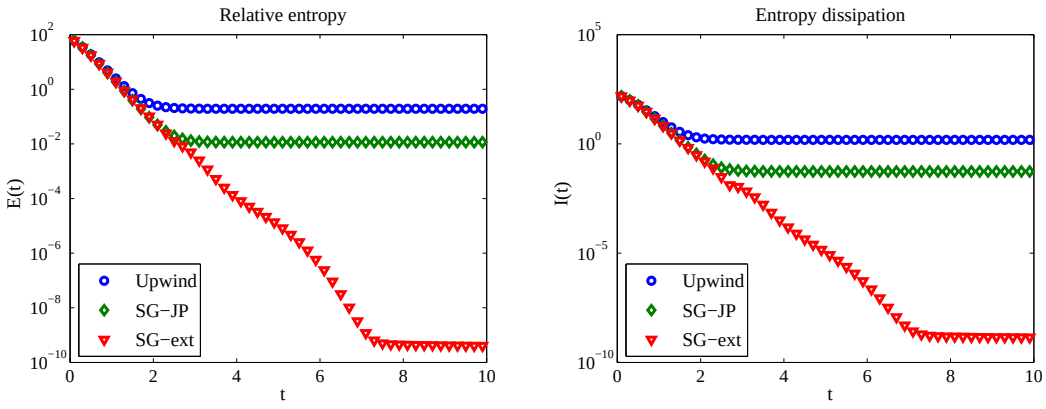


Figure 3.6: Evolution of the relative entropy $\mathcal{E}^n$ and its dissipation $\mathcal{I}^n$ in log-scale for different schemes.

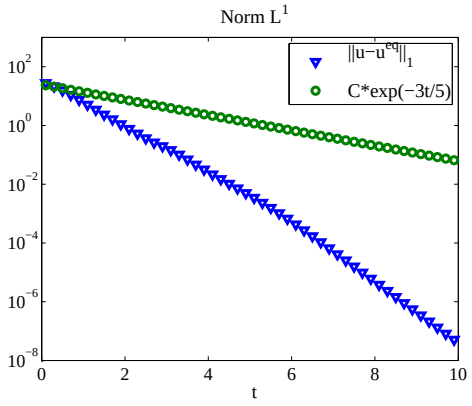


Figure 3.7: Decay rate of $\|U - U^{eq}\|_{L^1}$.

scheme, in such way that ensures that a particular type of steady-state is preserved. Moreover, this new scheme is easier to implement than existing schemes preserving steady-state. In addition, we have shown that there is convergence of our scheme in the nondegenerate case. The proof of this convergence is essentially based on a discrete $L^2\left(0,T;H^1\right)$ estimate (3.3.4). A first step to then prove the convergence in the degenerate case would be to show this estimate without using the uniform lower bound of u_δ.

Finally, we have observed that this scheme appears to be more accurate than the upwind one, even in the degenerate case. Indeed, we have applied it to the drift-diffusion model for semiconductors as well as to the porous media equation. In these two specific cases, we clearly underlined the efficiency of our scheme in order to preserve long-time behavior of the solutions. At this point, it still remains to prove rigorously this asymptotic behavior, by showing a similar estimate to the one of the continuous framework (3.1.5) for discrete energy and discrete dissipation.

Chapitre 4

Un schéma volumes finis d'ordre deux pour des équations paraboliques non linéaires dégénérées *

Dans ce chapitre, nous construisons un schéma volumes finis d'ordre deux en espace, pour des équations paraboliques non linéaires dégénérées qui admettent une fonctionnelle d'entropie. Pour nombre de modèles décrits par de telles équations (équation des milieux poreux, système de dérive-diffusion pour les semi-conducteurs,...), la solution du problème évolutif converge en temps long vers un état stationnaire. Le schéma présenté dans ce chapitre préserve les états stationnaires et fournit un comportement en temps long satisfaisant. De plus, il reste valide et précis à l'ordre deux en espace même dans le cas dégénéré. Après avoir décrit la construction de ce schéma numérique, nous présentons un certain nombre de résultats numériques qui d'une part mettent en évidence l'ordre élevé, à la fois dans les régimes non dégénérés et dégénérés, et d'autre part soulignent l'efficacité du schéma pour préserver l'asymptotique en temps long.

∗. Ce chapitre est un article écrit en collaboration avec F. Filbet, *A finite volume scheme for nonlinear degenerate parabolic equations* [19], accepté pour publication dans SIAM Journal on Scientific Computing.

Contents

4.1 Introduction

In this chapter we propose a second order accurate finite volume scheme for solving the following nonlinear, possibly degenerate parabolic equation: for $u : \mathbb{R}^+ \times \Omega \mapsto \mathbb{R}^+$ solution to

$$\begin{cases} \partial_t u = \operatorname{div}\left(f(u)\nabla V(x) + \nabla r(u)\right), \quad x \in \Omega, \quad t > 0, \\ \\ u(t=0,x) = u_0(x), \end{cases} \tag{4.1.1}$$

where $\Omega \subset \mathbb{R}^d$ is an open bounded domain or the whole space $\mathbb{R}^d$, $u \geq 0$ is a time-dependent density, f is a given function and $r \in \mathcal{C}^1(\mathbb{R}_+)$ is such that $r'(u) \geq 0$ and $r'(u)$ can vanish for certain values of u. Moreover, we assume that r and f are such that there exists a function h such that $r'(s) = h'(s)\, f(s)$, and that $f(u) \geq 0$. This assumption means that the problem we consider has a structure corresponding to an energy or entropy, or more generally a Lyapunov functional. Or aim is to design a scheme which preserves this physical property. Indeed, the equation (4.1.1) can be now written as

$$\partial_t u = \operatorname{div}\left(f(u)\,\nabla\left(V(x) + h(u)\right)\right), \tag{4.1.2}$$

and this equation admits an entropy functional: if we multiply (4.1.2) by $(V + h(u))$ and integrate over Ω, it yields

$$\frac{dE(t)}{dt} = -\mathcal{I}(t) \leq 0,$$

where the entropy E is defined by

$$E(t) := \int_\Omega \left(V\, u + H(u)\right)\, dx$$

with H an antiderivative of the function h, and the entropy dissipation $\mathcal{I}$ is given by

$$\mathcal{I}(t) = \int_\Omega f(u)\, \left|\nabla\left(V + h(u)\right)\right|^2\, dx.$$

A large variety of numerical methods have been proposed for the discretization of nonlinear degenerate parabolic equations: piecewise linear finite elements [13, 77, 115, 145, 146], cell-centered finite volume schemes [85, 87], vertex-centered finite volume schemes [147], finite difference methods [126], mixed finite element methods [7], local discontinuous Galerkin finite element methods [166], combined finite volume-finite element approach [88]. Schemes based on discrete BGK models have been proposed in [8], as well as characteristics-based methods considered in [56, 123]. Other approaches are either based on a suitable splitting technique [80], or based on the maximum principle and on perturbation and regularization [152]. Also high order schemes have been developed in [49, 135, 130], which is a crucial step getting an accurate approximation of the transient solution.
In this chapter our aim is to construct a second-order finite volume scheme preserving steady-states in order to obtain a satisfying long-time behavior for numerical solutions. Indeed, it has been observed in Chapter 3 that numerical schemes based on the preservation of steady states for degenerate parabolic problems offer a very accurate behavior of the approximate solution as time goes to infinity. To our knowledge, only few papers investigate this large-time asymptotic of numerical solutions. L. Gosse and G. Toscani proposed in [99] a scheme based on a formulation using the pseudo-inverse of the density's repartition function for porous media equation and fast-diffusion equation, and analysed the long-time behavior of approximate solutions. C. Chainais-Hillairet and F. Filbet studied in [51] a finite volume discretization for nonlinear drift-diffusion system and proved that the numerical solution converges to a steady-state when time goes to infinity. In [38], M. Burger, J. A. Carrillo and M. T. Wolfram proposed a mixed finite element method for nonlinear diffusion equations and proved convergence towards the steady-state in case of a nonlinear Fokker-Planck equation with uniformly convex potential. Here we propose a general way for designing a high-order scheme for nonlinear degenerate parabolic equations (4.1.1) admitting an entropy functional. This scheme preserves steady-states and entropy decay like those proposed in [16, 51, 99]. Moreover, it appears that a loss of accuracy can happen when the problem degenerates, causing a deterioration of the long-time behavior of the approximate solution. Our new scheme tackles this issue since it remains second-order accurate in space both in degenerate and non-degenerate regimes.

Before describing our numerical scheme, let us emphasize that for some models described by equation (4.1.1), the large-time asymptotic has been studied using entropy/entropy-dissipation arguments, which will be the starting point of our approach. On the one hand equation (4.1.1) with linear convection, namely $f(u) = u$, has been analysed by J. A. Carrillo, A. Jüngel, P. A. Markowich, G. Toscani and A. Unterreiter in [45]. On the other hand for equation (4.1.1) with nonlinear convection and linear diffusion a particular case has been studied in [47, 46, 160] by J. A. Carrillo, P. Laurençot, J. Rosado, F. Salvarani and G. Toscani. We will now remind some of the useful results contained in these papers.

Case of a linear convection. The paper [45] focuses on the long time asymptotic with exponential decay rate for

$$\partial_t u = \text{div}\,\left(u\nabla V(x) + \nabla r(u)\right), \quad x \in \Omega, \quad t > 0, \tag{4.1.3}$$

with initial condition $u(t=0,x) = u_0(x) \geq 0$, $u_0 \in L^1(\Omega)$ and

$$\int_\Omega u_0(x)\,dx =: M.$$

Equation (4.1.3) is supplemented either by a decay condition when $|x| \to \infty$ if $\Omega = \mathbb{R}^d$ or by a zero out-flux condition on $\partial\Omega$ if Ω is bounded. In the following, we assume that $r : \mathbb{R}_+ \to \mathbb{R}$ belongs to $\mathcal{C}^2(\mathbb{R}_+)$, is increasing and verifies $r(0) = 0$. We define

$$h(s) := \int_1^s \frac{r'(\tau)}{\tau}\,d\tau, \quad s \in (0,\infty), \tag{4.1.4}$$

and assume that $h \in L^1_{loc}\left([0,\infty)\right)$. Then

$$H(s) := \int_0^s h(\tau)\,d\tau, \quad s \in [0,\infty),$$

is well-defined, and $H'(s) = h(s)$ for all $s \geq 0$.
To analyze the large-time behavior to (4.1.3), stationary solutions u^{eq} of (4.1.3) in Ω are first studied:

$$u^{eq}\nabla V(x) + \nabla r(u^{eq}) = 0, \quad \int_\Omega u^{eq}(x)\,dx = M.$$

By using the definition (4.1.4) of h, this can be written as

$$u^{eq}\left(\nabla V(x) + \nabla h(u^{eq})\right) = 0, \quad \int_\Omega u^{eq}(x)\,dx = M,$$

and if $u^{eq} > 0$ in Ω, then one obtains

$$V(x) + h\left(u^{eq}(x)\right) = C \quad \forall x \in \Omega,$$

for some $C \in \mathbb{R}$. By considering the entropy functional

$$E(u) := \int_\Omega \left(V(x)\,u(x) \,+\, H(u(x))\right)\,dx,$$

a function $u^{eq,M} \in L^1(\Omega)$ is an equilibrium solution of (4.1.3) if and only if it is a minimizer of E in

$$\mathcal{C} \,=\, \left\{u \in L^1(\Omega),\ \int_\Omega u(x)\,dx = M\right\}.$$

Under some regularity assumptions on V, existence and uniqueness of an equilibrium solution is proved. Therefore, the long time behavior is investigated and the exponential decay of the relative entropy

$$\mathcal{E}\left(t\right) := E\left(u(t)\right) - E(u^{eq,M}) \tag{4.1.5}$$

is shown, using the exponential decay of the entropy dissipation

$$\mathcal{I}(t) := -\frac{d\mathcal{E}(t)}{dt} = \int_\Omega u(t,x)\,|\nabla\,(V(x) + h(u(t,x)))|^2\,dx.$$

Finally using a generalized Csiszar-Kullback inequality, it is proved that the solution $u(t,x)$ of (4.1.3) with $r(s) = \log(s)$ or $r(s) = s^m$, $m \geq 0$, converges to the equilibrium $u^{eq,M}(x)$ as $t \to \infty$ at an exponential rate.

Equation (4.1.3) includes many well-known equations governing physical phenomena as porous media or drift-diffusion models for semiconductors.

Example (the porous media equation). In the case $V(x) = |x|^2/2$ and $r(u) = u^m$, with $m > 1$, equation (4.1.3) is the porous media equation, which describes the flow of a gas through a porous interface. J. A. Carrillo and G. Toscani have proved in [48] that the unique stationary solution of the porous media equation is given by Barenblatt-Pattle type formula

$$u^{eq}(x) = \left(C_1 - \frac{m-1}{2m}|x|^2\right)_+^{1/(m-1)}, \tag{4.1.6}$$

where C_1 is a constant such that u^{eq} has the same mass as the initial data u_0. Moreover, the convergence of the solution $u(t,x)$ of the porous media equation to the Barenblatt-Pattle solution $u^{eq}(x)$ as $t \to \infty$ has been proved in [48], using the entropy method.

Example (the drift-diffusion model for semiconductors). The drift-diffusion model can also be interpreted in the formalism of (4.1.3). It is written as

$$\begin{cases} \partial_t N - \nabla\cdot(\nabla r(N) - N\nabla V) = 0, \\ \partial_t P - \nabla\cdot(\nabla r(P) + P\nabla V) = 0, \\ \Delta V = N - P - C, \end{cases} \tag{4.1.7}$$

where the unknowns are N the electron density, P the hole density and V the electrostatic potential, and C is the prescribed doping profile. The two continuity equations on the densities N and P correspond to (4.1.3) with $r(s) = s^\gamma$ the pressure function. These equations are supplemented with initial conditions $N_0(x)$ and $P_0(x)$ and physically motivated boundary conditions: Dirichlet boundary conditions $\overline{N}$, $\overline{P}$ and $\overline{V}$ on ohmic contacts Γ^D and homogeneous Neumann boundary conditions on insulating boundary segments Γ^N. The stationary drift-diffusion system admits a solution (N^{eq}, P^{eq}, V^{eq}) (see [138]), which is unique if in addition:

$$h(N^{eq}) - V^{eq} \begin{cases} = \alpha_N & \text{if} \quad N^{eq} > 0 \\ \geq \alpha_N & \text{if} \quad N^{eq} = 0 \end{cases}, \quad h(P^{eq}) + V^{eq} \begin{cases} = \alpha_P & \text{if} \quad P^{eq} > 0 \\ \geq \alpha_P & \text{if} \quad P^{eq} = 0 \end{cases}, \tag{4.1.8}$$

holds, and if the Dirichlet boundary conditions satisfy (4.1.8) and the compatibility condition (if $\overline{N}\,\overline{P} > 0$)

$$h(\overline{N}) + h(\overline{P}) = \alpha_N + \alpha_P. \tag{4.1.9}$$

In this case the thermal equilibrium (N^{eq}, P^{eq}, V^{eq}) is defined by

$$\text{(4.1.10)} \qquad \begin{cases} \Delta V^{eq} = g\left(\alpha_N + V^{eq}\right) - g\left(\alpha_P - V^{eq}\right) - C & \text{on } \Omega, \\ \\ N^{eq} = g\left(\alpha_N + V^{eq}\right), \;\; P^{eq} = g\left(\alpha_P - V^{eq}\right) & \text{on } \Omega, \end{cases}$$

where g is the generalized inverse of h, namely

$$g(s) = \begin{cases} h^{-1}(s) & \text{if} \quad h(0_+) < s < \infty, \\ 0 & \text{if} \quad s \leq h(0_+). \end{cases}$$

In the linear case $r(u) = u$, it has been proved by H. Gajewski and K. Gärtner in [92] that the solution to the transient system (4.1.7) converges to the thermal equilibrium state as $t \to \infty$ if the boundary conditions are in thermal equilibrium. A. Jüngel extends this result to a degenerate model with nonlinear diffusion in [119]. In both cases the key-point of the proof is an energy estimate with the control of the energy dissipation.

Case of a nonlinear convection. In [47, 46, 160], a nonlinear Fokker-Planck type equation modelling the relaxation of fermion and boson gases is studied. This equation corresponds to (4.1.1) with linear diffusion and nonlinear convection:

$$\text{(4.1.11)} \qquad \partial_t u = \text{div}\left(xu(1+ku) + \nabla u\right), \quad x \in \mathbb{R}^d, \quad t > 0,$$

with $k = 1$ in the boson case and $k = -1$ in the fermion case. The long-time asymptotic of this model has been studied in 1D for both cases [47], in any dimension for fermions [46] and in 3D for bosons [160]. The stationary solution of (4.1.11) is given by the Fermi-Dirac ($k = -1$) and Bose-Einstein ($k = 1$) distributions:

$$\text{(4.1.12)} \qquad u^{eq}(x) = \frac{1}{\beta e^{\frac{|x|^2}{2}} - k},$$

where $\beta \geq 0$ is such that u^{eq} has the same mass as the initial data u_0. The entropy functional is given by

$$E(u) := \int_{\mathbb{R}^d} \left(\frac{|x|^2}{2} u + u \log(u) - k(1+ku)\log(1+ku) \right) dx,$$

and the entropy dissipation is defined by

$$\mathcal{I}(t) := -\frac{d\mathcal{E}(t)}{dt} = \int_{\mathbb{R}^d} u(1+ku) \left| \nabla \left(\frac{|x|^2}{2} + \log\left(\frac{u}{1+ku} \right) \right) \right|^2 dx.$$

Then decay rates towards equilibrium are given in [47, 46] for fermion case in any dimension and for 1D boson case by relating the entropy and its dissipation. As in the case of a linear diffusion, the key-point of the proof is an entropy estimate with the control of its dissipation.

Concerning 3D boson case, it is proved in [160] that for sufficiently large initial mass, the solution blows up in finite time.
Let us also mention that a more general class of Fokker-Planck type equations for bosons with linear diffusion and super-linear drift is studied in [15]:

$$\partial_t u = \mathrm{div}(xu(1+u^N) + \nabla u), \tag{4.1.13}$$

where $N > 0$ is a given constant. For $N > 2$, there is a phenomenon of critical mass in dimension 1. It is proved by minimizing an entropy functional that starting from an initial distribution with a super-critical mass, the solution develops a singular part localized in the origin.

As explained above, it has been proved by entropy/entropy dissipation techniques that the solution to (4.1.1) converges to a steady-state as time goes to infinity often with an exponential time decay rate. Our aim is to propose a numerical scheme considering these problems and for which we can obtain a discrete entropy estimate as in the continuous case. In [11, 43, 38] temporal semi-discretizations have been proposed and semi-discrete entropy estimates have been proved. However, when the problem is spatially discretized a saturation of the entropy and its dissipation may appear, due to the spatial discretization error. This emphasizes the importance of considering spatial discretization techniques which preserve the steady-states and the entropy dissipation. This point of view has been already adopted in [16, 51] but both schemes do not provide really satisfying results when the equation degenerates. Indeed both schemes degenerate in the upwind flux if the diffusion vanishes and then are only first order accurate in space. Thus we propose in this chapter a finite volume scheme for nonlinear parabolic equations, possibly degenerate, possessing an entropy functional. We focus on the spatial discretization, with a twofold objective. On the one hand we require preserving steady-states in order to obtain a satisfying long-time behavior of the approximate solution. On the other hand the scheme proposed remains valid and second order accurate in space even in the degenerate case. The main idea of our new scheme is to discretize together the convective and diffusive parts of the equation (4.1.1) to obtain a flux which preserves equilibrium and to use a slope-limiter method to get second-order accuracy even in the degenerate case.

The plan of the chapter is as follows. In Section 4.2, we construct the finite volume scheme. We first focus on the case of a linear diffusion (4.1.3). Then we extend this construction to the general case (4.1.1). In Section 4.3 we give some basic properties of the scheme and a semidiscrete entropy estimate for the case of a linear diffusion (4.1.3). We end in Section 4.4 by presenting some numerical results. We first verify experimentally the second order accuracy in space of our scheme, even in the degenerate case. Then we focus on the long-time behavior. The scheme is applied to the physical models introduced above and the numerical results confirm its efficiency to preserve the large-time asymptotics. Finally we propose a test case with both nonlinear convection and diffusion.

4.2 Presentation of the numerical scheme

In this section we present our new finite volume scheme for (4.1.1). For simplicity purposes, we consider the problem in one space dimension. It will be straightforward to generalize this construction for Cartesian meshes in multidimensional case.
In a one-dimensional setting, $\Omega = (a,b)$ is an interval of $\mathbb{R}$. We consider a mesh for the domain (a,b), which is not necessarily uniform *i.e.* a family of N_x control volumes $(K_i)_{i=1,...,N_x}$ such that $K_i = \left]x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right[$ with $x_i = (x_{i-\frac{1}{2}} + x_{i+\frac{1}{2}})/2$ and

$$a = x_{\frac{1}{2}} < x_1 < x_{\frac{3}{2}} < ... < x_{i-\frac{1}{2}} < x_i < x_{i+\frac{1}{2}} < ... < x_{N_x} < x_{N_x+\frac{1}{2}} = b.$$

Let us set

$$\begin{aligned} \Delta x_i &= x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}, \quad \text{for } 1 \le i \le N_x, \\ \Delta x_{i+\frac{1}{2}} &= x_{i+1} - x_i, \quad \text{for } 1 \le i \le N_x - 1. \end{aligned}$$

Let Δt be the time step. We set $t^n = n\Delta t$. A time discretization of $(0,T)$ is then given by the integer value $N_T = E(T/\Delta t)$ and by the increasing sequence of $(t^n)_{0\le n\le N_T}$.
First of all, the initial condition is discretized on each cell K_i by:

$$U_i^0 = \frac{1}{\Delta x_i}\int_{K_i} u_0(x)\,dx, \quad i = 1,...L.$$

The finite volume scheme is obtained by integrating the equation (4.1.1) over each control volume K_i and over each time step. Concerning the time discretization, we can choose any explicit method (forward Euler, Runge-Kutta,...). Since in this chapter we are interested in the spatial discretization, we will only consider a forward Euler method afterwards. Let us now focus on the spatial discretization.
We denote by $U_i(t)$ an approximation of the mean value of u over the cell K_i at time t. By integrating the equation (4.1.1) on K_i, we obtain the semi-discrete numerical scheme:

$$\Delta x_i \frac{d}{dt}U_i + \mathcal{F}_{i+\frac{1}{2}} - \mathcal{F}_{i-\frac{1}{2}} = 0, \tag{4.2.1}$$

where $\mathcal{F}_{i+\frac{1}{2}}$ is an approximation of the flux $-[f(u)\partial_x V + \partial_x r(u)]$ at the interface $x_{i+\frac{1}{2}}$ which remains to be defined.

Case of a linear convection ($f(u) = u$). To explain our approach we first define the numerical flux for equation (4.1.3). The main idea is to discretize together the convective and the diffusive parts. To this end, we write $[u\partial_x V + \partial_x r(u)]$ as $u\left[\partial_x\left(V + h(u)\right)\right]$, where h is defined by (4.1.4). Then we will consider $-\partial_x\left(V + h(u)\right)$ as a velocity and denote by $A_{i+\frac{1}{2}}$ an approximation of this velocity at the interface $x_{i+\frac{1}{2}}$:

$$A_{i+\frac{1}{2}} = -dV_{i+\frac{1}{2}} - dh(U)_{i+\frac{1}{2}},$$

where $dV_{i+\frac{1}{2}}$ and $dh(U)_{i+\frac{1}{2}}$ are centered approximations of $\partial_x V$ and $\partial_x h(u)$ respectively, namely

$$dV_{i+\frac{1}{2}} = \frac{V(x_{i+1}) - V(x_i)}{\Delta x_{i+\frac{1}{2}}}, \quad dh(U)_{i+\frac{1}{2}} = \frac{h(U_{i+1}) - h(U_i)}{\Delta x_{i+\frac{1}{2}}}.$$

Now we apply the standard upwind method and then define our new numerical flux, called fully upwind flux, as

$$\mathcal{F}_{i+\frac{1}{2}} = F(U_i, U_{i+1}) = A^+_{i+\frac{1}{2}} U_i - A^-_{i+\frac{1}{2}} U_{i+1}, \tag{4.2.2}$$

where $x^+ = \max(0, x)$ and $x^- = \max(0, -x)$. This method is only first-order accurate. To obtain second-order accuracy, we replace in (4.2.2) U_i and U_{i+1} by $U_{i+\frac{1}{2},-}$ and $U_{i+\frac{1}{2},+}$ respectively, which are reconstructions of u at the interface defined by:

$$\begin{cases} U_{i+\frac{1}{2},-} \;=\; U_i + \frac{1}{2}\phi\left(\theta_i\right)\left(U_{i+1} - U_i\right), \\ \\ U_{i+\frac{1}{2},+} \;=\; U_{i+1} - \frac{1}{2}\phi\left(\theta_{i+1}\right)\left(U_{i+2} - U_{i+1}\right), \end{cases} \tag{4.2.3}$$

with

$$\theta_i = \frac{U_i - U_{i-1}}{U_{i+1} - U_i}$$

and ϕ is a slope-limiter function (setting $\phi = 0$ gives the classical upwind flux). From now on we will consider the second-order fully upwind scheme defined with the Van Leer limiter:

$$\phi(\theta) = \frac{\theta + |\theta|}{1 + |\theta|}.$$

General case. We now consider the general case where both diffusion and convection are nonlinear in (4.1.1). We assume that $f(u) \geq 0$ and that we can define $h(u)$ such that $h'(u)f(u) = r'(u)$. Then the equation (4.1.1) admits an entropy, as explained in the introduction. Following the same idea as above, we use the following expression of the flux

$$f(u)\partial_x V + \partial_x r(u) \;=\; \partial_x\left(V + h(u)\right) f(u), \tag{4.2.4}$$

and define the numerical flux as a local Lax-Friedrichs:

$$\mathcal{F}_{i+\frac{1}{2}} = \frac{A_{i+\frac{1}{2}}}{2}\left(f(U_i) + f(U_{i+1})\right) - \frac{\left|A_{i+\frac{1}{2}}\right| \alpha_{i+\frac{1}{2}}}{2}\left(U_{i+1} - U_i\right), \tag{4.2.5}$$

where

$$A_{i+\frac{1}{2}} = -dV_{i+\frac{1}{2}} - dh(U)_{i+\frac{1}{2}},$$

and

$$\alpha_{i+\frac{1}{2}} = \max\left(\left|f'(u)\right|\right) \text{ over all } u \text{ between } U_i \text{ and } U_{i+1}.$$

As above, we replace U_i and U_{i+1} in (4.2.5) by reconstructions $U_{i+\frac{1}{2},-}$ and $U_{i+\frac{1}{2},+}$ defined by (4.2.3) to obtain a second-order scheme.

We can now summarize our new numerical flux by:

(4.2.6)
$$\begin{cases} \mathcal{F}_{i+\frac{1}{2}} = \dfrac{A_{i+\frac{1}{2}}}{2}\left(f(U_{i+\frac{1}{2},-}) + f(U_{i+\frac{1}{2},+})\right) - \dfrac{\left|A_{i+\frac{1}{2}}\right|\alpha_{i+\frac{1}{2}}}{2}\left(U_{i+\frac{1}{2},+} - U_{i+\frac{1}{2},-}\right), \\ A_{i+\frac{1}{2}} = -dV_{i+\frac{1}{2}} - dh(U)_{i+\frac{1}{2}}, \\ \alpha_{i+\frac{1}{2}} = \max\left(\left|f'(u)\right|\right) \text{ over all } u \text{ between } U_i \text{ and } U_{i+1}, \\ U_{i+\frac{1}{2},-} = U_i + \dfrac{1}{2}\phi\left(\theta_i\right)\left(U_{i+1} - U_i\right), \\ U_{i+\frac{1}{2},+} = U_{i+1} - \dfrac{1}{2}\phi\left(\theta_{i+1}\right)\left(U_{i+2} - U_{i+1}\right), \end{cases}$$

where either a first-order scheme

$$\phi(\theta) = 0, \tag{4.2.7}$$

or a second order scheme

$$\phi(\theta) = \frac{\theta + |\theta|}{1 + |\theta|}. \tag{4.2.8}$$

Remark 1 (Generalization to multidimensional case). It is straightforward to define the scheme for Cartesian meshes in multidimensional case: the 1D formula can be used as it is in any of the Cartesian directions. However, the construction of the scheme on unstructured meshes is more complicated. More precisely, it is easy to define the first order scheme on such grids, but the difficulty is to obtain high-order accuracy. As in the one dimensional case, the idea is to replace the first-order flux $F(U_i, U_j)$, where U_i, U_j are the constant values on each side of an edge $\Gamma_{ij} = K_i \cap K_j$, by $F(U_{ij}, U_{ji})$, where U_{ij}, U_{ji} are second-order approximations of the solution on each side of the edge Γ_{ij}. More precisely, we need to obtain piecewise linear functions on each triangle instead of piecewise constant functions. For more details concerning these questions, see for example [76, 98] and the references therein.

4.3 Properties of the scheme

In this section, we present some important properties of the scheme. We would like to emphasize here the preservation of the equilibrium and the entropy estimate, which are two crucial properties to study the scheme. Concerning a more advanced analysis of the scheme, we can apply the same techniques as in Chapter 3, but this is not our purpose here.

4.3.1 The semi-discrete scheme

In this subsection, we study the semi-discrete scheme (4.2.1)-(4.2.6)-(4.2.8) and consider the equation (4.1.3) on a bounded domain with homogeneous Neumann boundary conditions. We assume that $r \in \mathcal{C}^1(\mathbb{R}_+)$ is strictly increasing and h is defined by (4.1.4). Then we consider a primitive H of h, which is strictly convex since r is strictly increasing. We denote by $(U_i^{eq})_{i=1,...,N_x}$ an approximation of the equilibrium solution u^{eq}. This approximation verifies

$$dh\,(U^{eq})_{i+\frac{1}{2}} + dV_{i+\frac{1}{2}} = 0 \quad \forall i = 0, ..., N_x, \tag{4.3.1}$$

and

$$\sum_{i=1}^{N_x} \Delta x_i U_i^{eq} = \sum_{i=1}^{N_x} \Delta x_i U_i^0 =: \overline{M}.$$

A semi discrete version of the relative entropy $\mathcal{E}$ defined by (4.1.5) is given by

$$\mathcal{E}_\Delta(t) := \sum_{i=1}^{N_x} \Delta x_i \big(H\,(U_i(t)) - H\,(U_i^{eq}) - h\,(U_i^{eq})\,(U_i(t) - U_i^{eq})\,\big). \tag{4.3.2}$$

We also introduce the semi discrete version of the entropy dissipation

$$\mathcal{I}_\Delta(t) := \sum_{i=0}^{N_x} \Delta x_{i+\frac{1}{2}} \left|A_{i+\frac{1}{2}}\right|^2 \min\left(U_{i+\frac{1}{2},-}(t), U_{i+\frac{1}{2},+}(t)\right).$$

Proposition 4.3.1. *Assume that the initial data $U_i(0)$ is nonnegative. Then, the finite volume scheme (4.2.1)-(4.2.6)-(4.2.8) for equation (4.1.3) satisfies*

(i) the preservation of the nonnegativity of $U_i(t)$,

(ii) the preservation of the equilibrium,

(iii) the entropy estimate: for $0 < t_1 \le t_2 < \infty$,

$$0 \le \mathcal{E}_\Delta(t_2) + \int_{t_1}^{t_2} \mathcal{I}_\Delta(t)\,dt \le \mathcal{E}_\Delta(t_1).$$

Proof. To prove the preservation of nonnegativity, we need to check that

$$F\left(U_{i+\frac{1}{2},-}, U_{i+\frac{1}{2},+}\right) - F\left(U_{i-\frac{1}{2},-}, U_{i-\frac{1}{2},+}\right) \le 0 \tag{4.3.3}$$

whenever $U_i = 0$.
When $U_i = 0$, we have $U_i \le U_{i+1}$ and $U_i \le U_{i-1}$, and then $\theta_i \le 0$, which gives $\phi(\theta_i) = 0$ and finally

$$U_{i+\frac{1}{2},-} = U_{i-\frac{1}{2},+} = U_i = 0.$$

Then we get

$$F\left(U_{i+\frac{1}{2},-}, U_{i+\frac{1}{2},+}\right) - F\left(U_{i-\frac{1}{2},-}, U_{i-\frac{1}{2},+}\right) = -A^-_{i+\frac{1}{2}} U_{i+\frac{1}{2},+} - A^+_{i-\frac{1}{2}} U_{i-\frac{1}{2},-}.$$

Moreover, $U_{i-\frac{1}{2},-}$ is given by

$$U_{i-\frac{1}{2},-} \;=\; \left(1-\frac{\phi(\theta_{i-1})}{2}\right) U_{i-1},$$

which is nonnegative since $\phi(\theta) \leq 2$ for all θ.

On the other hand, we deal with $U_{i+\frac{1}{2},+}$, and get that either $\theta_{i+1} \leq 0$, then $U_{i+\frac{1}{2},+} = U_{i+1} \geq 0$, or we have $\theta_{i+1} > 0$, that is $U_{i+2} \geq U_{i+1}$ and since $\phi(\theta) \leq 2\theta$ for all $\theta \geq 0$, we get

$$U_{i+\frac{1}{2},+} \geq U_{i+1} - \theta_{i+1}\left(U_{i+2} - U_{i+1}\right) = U_{i+1} - \left(U_{i+1} - U_i\right) = 0.$$

We conclude that (4.3.3) always holds when $U_i = 0$, which gives (i).
The part (ii) is clear by construction: at the equilibrium, we have $dh(U)_{i+\frac{1}{2}} + dV_{i+\frac{1}{2}} = 0$, which is exactly $A_{i+\frac{1}{2}} = 0$ and then $\mathcal{F}_{i+\frac{1}{2}} = 0$.
By definition (4.3.2) of $\mathcal{E}_\Delta(t)$ and since $H'(s) = h(s)$ for all $s \geq 0$, we have

$$\frac{d\mathcal{E}_\Delta}{dt}(t) = \sum_{i=1}^{N_x} \Delta x_i \left(h((U_i(t)) - h(U_i^{eq})\right) \frac{dU_i}{dt}(t).$$

Using the numerical scheme (4.2.1), we get

$$\frac{d\mathcal{E}_\Delta}{dt}(t) = -\sum_{i=1}^{N_x} \left(h((U_i(t)) - h(U_i^{eq})\right) \left(\mathcal{F}_{i+\frac{1}{2}} - \mathcal{F}_{i-\frac{1}{2}}\right),$$

and then a discrete integration by parts yields (using the homogeneous Neumann boundary conditions)

$$\frac{d\mathcal{E}_\Delta}{dt}(t) = \sum_{i=0}^{N_x} \Delta x_{i+\frac{1}{2}} \left(dh(U(t))_{i+\frac{1}{2}} - dh(U^{eq})_{i+\frac{1}{2}}\right) \mathcal{F}_{i+\frac{1}{2}}.$$

Since by (4.3.1) we have $dh\left(U^{eq}\right)_{i+\frac{1}{2}} = -dV_{i+\frac{1}{2}}$, we obtain

$$\begin{aligned}
\frac{d\mathcal{E}_\Delta}{dt}(t) &= -\sum_{i=0}^{N_x} \Delta x_{i+\frac{1}{2}} A_{i+\frac{1}{2}} \left(A^+_{i+\frac{1}{2}} U_{i+\frac{1}{2},-}(t) - A^-_{i+\frac{1}{2}} U_{i+\frac{1}{2},+}(t)\right) \\
&\leq -\sum_{i=0}^{N_x} \Delta x_{i+\frac{1}{2}} \left|A_{i+\frac{1}{2}}\right|^2 \min\left(U_{i+\frac{1}{2},-}(t), U_{i+\frac{1}{2},+}(t)\right).
\end{aligned}$$

Finally we get (iii) by integrating between t_1 and t_2. □

4.3.2 The fully-discrete scheme

In this subsection we consider the fully-discrete scheme obtained by using the forward Euler method. We denote by U_i^n an approximation of the mean value of u over the cell K_i at time $t^n = n\Delta t$. The fully-discrete scheme is given by:

$$\mathrm{m}(K_i)\,\frac{U_i^{n+1} - U_i^n}{\Delta t} \;+\; \mathcal{F}^n_{i+\frac{1}{2}} - \mathcal{F}^n_{i-\frac{1}{2}} \;=\; 0, \tag{4.3.4}$$

where the numerical flux $\mathcal{F}_{i+\frac{1}{2}}$ is defined by (4.2.6)-(4.2.8).

Proposition 4.3.2. *For $n \geq 0$, assume that $U_i^n \geq 0$ for all $i = 1, ..., N_x$. Then under the CFL condition*

$$\Delta t \max_i \left| V(x_{i+1}) - V(x_i) - h(U_{i+1}^n) + h(U_i^n) \right| \ \leq \ \frac{1}{2} \min_i \Delta x_i^2, \tag{4.3.5}$$

the fully-discrete first-order scheme (4.2.6)-(4.2.7) and (4.3.4) for equation (4.1.3) preserves the nonnegativity of U_i, which means that $U_i^{n+1} \geq 0$ for all $i = 1, ..., N_x$, and the steady-states solution.

Proof. Using the definition (4.3.4)-(4.2.6)-(4.2.7) of the fully-discrete first-order scheme, we get for all $i = 1, ..., N_x$

$$\begin{aligned} U_i^{n+1} = & \left(1 - \frac{\Delta t}{\Delta x_i} \left(\left(A_{i+\frac{1}{2}}^n \right)^+ + \left(A_{i-\frac{1}{2}}^n \right)^- \right) \right) U_i^n \\ & + \frac{\Delta t}{\Delta x_i} \left(A_{i+\frac{1}{2}}^n \right)^- U_{i+1}^n + \frac{\Delta t}{\Delta x_i} \left(A_{i-\frac{1}{2}}^n \right)^+ U_{i-1}^n. \end{aligned}$$

Thus we deduce that $U_i^{n+1} \geq 0$ as soon as $\dfrac{\Delta t}{\Delta x_i} \left(\left(A_{i+\frac{1}{2}}^n \right)^+ + \left(A_{i-\frac{1}{2}}^n \right)^- \right) \leq 1$, which is necessarily the case from (4.3.5), using the definition of $A_{i+\frac{1}{2}}^n$. □

Remark 2. This result is not surprising since the stability condition for an explicit discretization of a parabolic equation requires the time step to be limited by a power two of the space step.

4.4 Numerical simulations

In this section, we present several numerical results performed by using our new fully-upwind flux. In all the numerical experiments performed, since our purpose is to focus on the spatial discretization, we choose a forward Euler method for the time discretization. As explained above, this choice of an explicit time discretization implies that the time step has to be limited by the square of the space step. Since we want to study the spatial accuracy of our scheme, we voluntarily choose a small time step in the first part of this section. Furthermore, since the CFL condition (4.3.5) becomes far less restrictive when the problem degenerates or when the solution tends to the equilibrium, we can use an adaptative time step. Nevertheless, a fully implicit scheme would be also suitable for the long time behavior of the numerical solution since in that case the numerical solution satisfies an entropy inequality, which is not the case with the explicit discretization we choose.

We first study the spatial order of convergence of the scheme for linear convection in both non degenerate and degenerate cases. Then we will apply it to the physical models presented in the introduction: the porous media equation, the drift-diffusion system for semiconductors and the nonlinear Fokker-Planck equation for bosons and fermions. The

results underline the efficiency of the scheme to preserve long-time behavior of the solutions. Finally we apply the scheme to a fully nonlinear problem: the Buckley-Leverett equation.
Below we make comparison between the finite volume schemes (4.3.4) defined with the following numerical fluxes:

– **The first-order fully upwind flux**, given by

$$(\mathbf{FU1}) \qquad \mathcal{F}_{i+\frac{1}{2}} = \frac{A_{i+\frac{1}{2}}}{2}\left(f(U_i)+f(U_{i+1})\right) - \frac{\left|A_{i+\frac{1}{2}}\right|\alpha_{i+\frac{1}{2}}}{2}\left(U_{i+1}-U_i\right),$$

with $A_{i+\frac{1}{2}}$, $\alpha_{i+\frac{1}{2}}$ defined in (4.2.6).

– **The second-order fully upwind flux**, given by

$$(\mathbf{FU2}) \quad \mathcal{F}_{i+\frac{1}{2}} = \frac{A_{i+\frac{1}{2}}}{2}\left(f(U_{i+\frac{1}{2},-}) + f(U_{i+\frac{1}{2},+})\right) - \frac{\left|A_{i+\frac{1}{2}}\right|\alpha_{i+\frac{1}{2}}}{2}\left(U_{i+\frac{1}{2},+} - U_{i+\frac{1}{2},-}\right).$$

– **The classical upwind flux,** introduced and studied in [85]. It is valid for linear convection and for both linear and nonlinear diffusion. The diffusion term is discretized classically by using a two-points flux and the convection term is discretized with the upwind flux. This flux has then been used for the drift-diffusion system for semiconductors [52, 53, 54]. It is defined for equation (4.1.3) by

$$(\mathbf{CU}) \qquad \mathcal{F}_{i+\frac{1}{2}} = \left(-dV_{i+\frac{1}{2}}\right)^{+} U_i - \left(-dV_{i+\frac{1}{2}}\right)^{-} U_{i+1} - \frac{r\left(U_{i+1}\right) - r\left(U_i\right)}{\Delta x_{i+\frac{1}{2}}}.$$

– **The Scharfetter-Gummel flux and its extension for nonlinear diffusion.** This scheme is widely used in the semiconductors framework in the case of a linear diffusion. It has been proposed in [114, 158] for the numerical approximation of the 1D drift-diffusion model. This scheme preserves equilibrium and is second-order accurate [132]. The definition of the Scharfetter-Gummel flux has been extended to the case of a nonlinear diffusion in Chapter 3. For equation (4.1.3) this flux is written

$$(\mathbf{SGext}) \quad \mathcal{F}_{i+\frac{1}{2}} = \frac{dr_{i+\frac{1}{2}}}{\Delta x_{i+\frac{1}{2}}}\left[B\left(\frac{\Delta x_{i+\frac{1}{2}} dV_{i+\frac{1}{2}}}{dr_{i+\frac{1}{2}}}\right)U_i - B\left(-\frac{\Delta x_{i+\frac{1}{2}} dV_{i+\frac{1}{2}}}{dr_{i+\frac{1}{2}}}\right)U_{i+1}\right],$$

where

$$\begin{cases} B(x) = \dfrac{x}{e^x - 1} \text{ for } x \neq 0, \quad B(0) = 1, \\ dr_{i+\frac{1}{2}} = dr\left(U_i, U_{i+1}\right), \end{cases}$$

with for a, $b \in \mathbb{R}_+$,

$$dr(a,b) = \begin{cases} \dfrac{h(b)-h(a)}{\log(b)-\log(a)} & \text{if } ab > 0 \text{ and } a \neq b, \\ r'\left(\dfrac{a+b}{2}\right) & \text{elsewhere.} \end{cases}$$

4.4.1 Order of convergence

In this subsection, we test the spatial accuracy of the scheme for linear convection ($f(s) = s$). We first consider a test case in 1D on $(0,T) \times (-1,1)$ with $\partial_x V = -1$. The time step is taken equal to $\Delta t = 10^{-8}$ to study the order of convergence with respect to the spatial step size. The boundary conditions are periodic. Since we don't know an exact solution of the problem, we compute relative errors. More precisely, an estimation of the relative error in L^1 norm at time T is given by

$$e_{2\Delta x} = \|u_{\Delta x}(T) - u_{2\Delta x}(T)\|_{L^1(\Omega)},$$

where $u_{\Delta x}$ represents the approximation computed from a mesh of size Δx. The numerical scheme is said to be k-th order if $e_{2\Delta x} \leq C\Delta x^k$, for all $0 < \Delta x \ll 1$.

Example 1 (Non degenerate case). We first take $r(s) = s^2$, thus $r'(0) = 0$ and $r'(s) > 0$ for all $s > 0$. The initial data is

$$u_0(x) = 0.5 + 0.5\sin(\pi x), \quad x \in (-1,1)$$

and the final time $T = 0.1$. In Figure 4.1, we represent the evolution of the approximate solution computed on a fine mesh made of 3200 cells, with the scheme (**FU2**). Since the solution becomes strictly positive for all $t > 0$, this problem is not degenerate. In

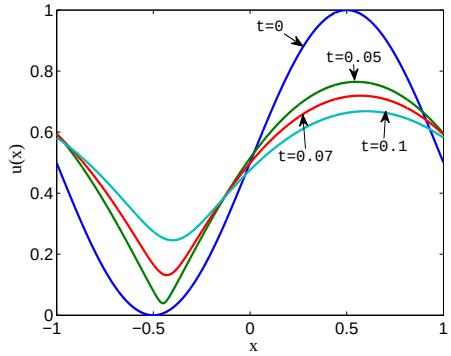


Figure 4.1: Example 1 - Evolution of the approximate solution computed on a fine mesh.

Table 4.1 we compare the order of convergence in L^1 norm of the Scharfetter-Gummel extended scheme (**SGext**) and of our first and second order fully upwind fluxes (**FU1**)-(**FU2**). It appears that the Scharfetter-Gummel scheme is second order accurate, as expected since the diffusion is not degenerate. Moreover, we verify experimentally that our scheme (**FU2**) is second-order accurate and we notice that the L^1 error obtained with it is smaller than that obtained with the Scharfetter-Gummel extended scheme.

N_x	L^1 error **SGext**	Order	L^1 error **FU1**	Order	L^1 error **FU2**	Order
100	$1.451.10^{-4}$	2	$2.667.10^{-3}$	0.87	$8.237.10^{-5}$	1.87
200	$3.619.10^{-5}$	2	$1.398.10^{-3}$	0.93	$2.208.10^{-5}$	1.9
400	$9.027.10^{-6}$	2	$7.156.10^{-4}$	0.97	$5.778.10^{-6}$	1.93
800	$2.251.10^{-6}$	2	$3.621.10^{-4}$	0.98	$1.485.10^{-6}$	1.96
1600	$5.614.10^{-7}$	2	$1.822.10^{-4}$	0.99	$3.772.10^{-7}$	1.98

Table 4.1: Example 1 - Experimental spatial order of convergence in L^1 norm.

Example 2 (Degenerate case). We still consider the same test case, but now with

$$r(s) = \begin{cases} (s-1)^3 & \text{if } s \geq 1, \\ 0 & \text{elsewhere,} \end{cases}$$

then $r'(s) = 0$ for all $s \in (0,1)$. The initial data is

$$u_0(x) = 1 + 0.5\sin(\pi x) \quad x \in (-1,1),$$

and the final time is $T = 0.01$. The diffusion vanishes in $\{x \in (-1,1) : u(x) \leq 1\}$, which is not empty, then this test case is degenerate. In Figure 4.2, we represent the evolution of the deviation from the initial data of the approximate solution computed on a fine mesh made of 3200 cells with the scheme (**FU2**). We observe a loss of regularity during the evolution. In Table 4.2 we compare the order of convergence in L^1 norm

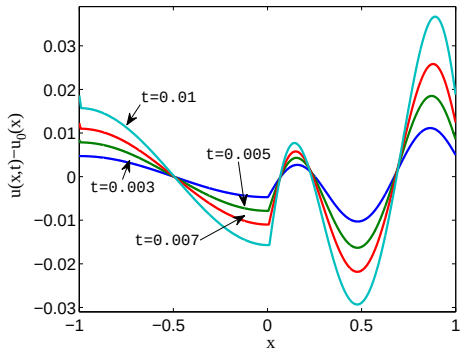


Figure 4.2: Example 2 - Evolution of the deviation from the initial data $u(t) - u_0$.

of the Scharfetter-Gummel extended scheme (**SGext**) and of our first and second order fully upwind fluxes (**FU1**)-(**FU2**). In this case where r' vanishes on a whole interval, it appears that the second-order scheme (**FU2**) is more accurate than the two others schemes. The Scharfetter-Gummel extended scheme is only one order accurate while second-order accuracy is almost preserved with our new scheme, in spite of the loss of regularity of the solution observed in Figure 4.2.

N_x	L^1 error **SGext**	Order	L^1 error **FU1**	Order	L^1 error **FU2**	Order
100	$3.074.10^{-4}$	0.96	$2.697.10^{-4}$	0.55	$1.053.10^{-4}$	1.83
200	$1.554.10^{-4}$	0.98	$1.531.10^{-4}$	0.82	$2.830.10^{-5}$	1.90
400	$7.834.10^{-5}$	0.99	$8.096.10^{-5}$	0.92	$8.040.10^{-6}$	1.82
800	$3.928.10^{-5}$	1	$4.163.10^{-5}$	0.96	$2.288.10^{-6}$	1.81
1600	$1.966.10^{-5}$	1	$2.111.10^{-5}$	0.98	$6.576.10^{-7}$	1.80

Table 4.2: Example 2 - Experimental spatial order of convergence in L^1 norm.

Example 3 (Degenerate case). Finally we consider the equation (4.1.3) on $(0,T)\times\Omega=(0,1/2)\times(0,1)$ with $r(s)=\max(u-1,0)$ and $\partial_x V=-1$. The initial data is $u_0(x)=0$ and we consider the following Dirichlet boundary conditions:

$$\left\{\begin{array}{rcl} u(t,0) &=& e^{2t} \\ u(t,1) &=& 0 \end{array}\right. , \quad t\in(0,T).$$

The exact solution is then given by

$$u(t,x)=\left\{\begin{array}{ll} \exp(2t-x) & \text{if} \quad x<2t, \\ 0 & \text{if} \quad x>2t. \end{array}\right.$$

We compute the solution up to $t=0.3$ with $\Delta t=10^{-4}$ and $N_x=40$ uniform cells. The results are shown in Figure 4.3. This example works well and it illustrates the advantage of using a high-order method even in the case of a discontinuous solution, since the shock is less diffused with our scheme (**FU2**) than with the three others.

4.4.2 The drift-diffusion system for semiconductors

We now consider the drift-diffusion system for semiconductors (4.1.7). In the two following examples, the Dirichlet boundary conditions satisfy (4.1.8)-(4.1.9), so the thermal equilibrium is uniquely defined by (4.1.10). We compute $(N_i^{eq},P_i^{eq},V_i^{eq})_{i=1,\dots,N_x}$ an approximation of this equilibrium with the finite volume scheme proposed by C. Chainais-Hillairet and F. Filbet in [51].

Example 4. Firstly we consider a 1D test case on $\Omega=(0,1)$. We take $r(s)=s^2$. Initial data are

$$N_0(x)=\left\{\begin{array}{lll} 0 & \text{for} & x\le 0.5 \\ 1 & \text{for} & x>0.5 \end{array}\right. , \quad P_0(x)=\left\{\begin{array}{lll} 1 & \text{for} & x\le 0.5 \\ 0 & \text{for} & x>0.5 \end{array}\right. ,$$

and we consider the following Dirichlet boundary conditions

$$\begin{array}{lll} N(0,t)=0, & P(0,t)=1, & V(0,t)=-1, \\ N(1,t)=1, & P(1,t)=0, & V(1,t)=1. \end{array}$$

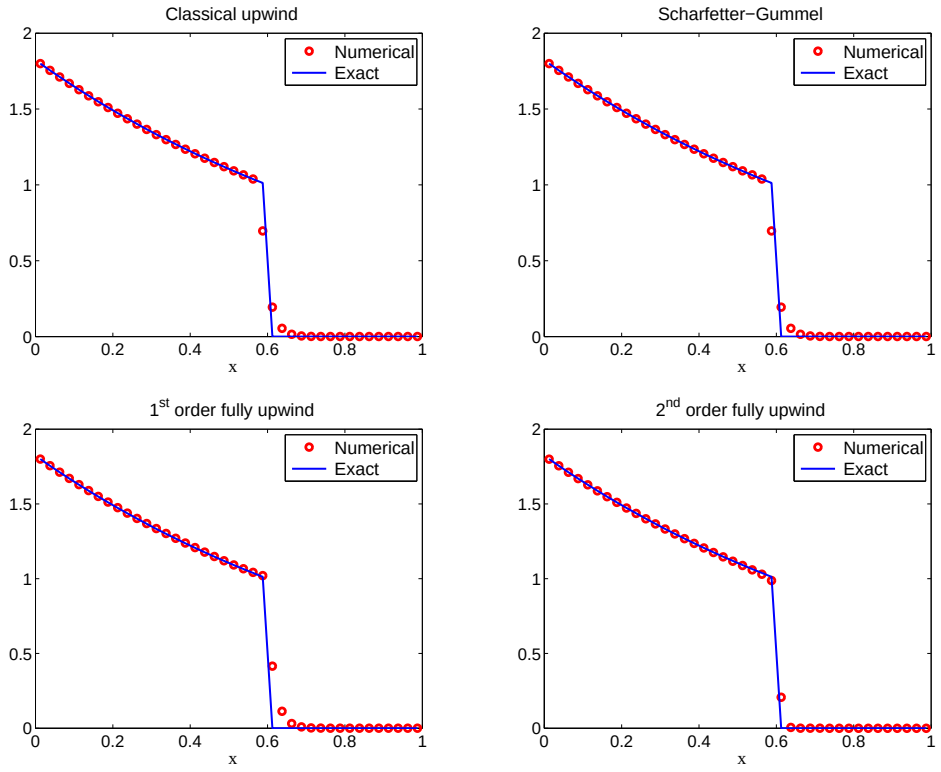


Figure 4.3: Example 3 - Numerical and exact solution computed at $t = 0.3$ with different schemes.

The doping profile is

$$C(x) = \begin{cases} -1 & \text{for} \quad x \leq 0.5, \\ +1 & \text{for} \quad x > 0.5. \end{cases}$$

The time step is $\Delta t = 5.10^{-5}$ and the final time $T = 10$. The domain $(0, 1)$ is divided into $N_x = 64$ uniform cells.
In Figure 4.4, we compare the discrete relative energy $\mathcal{E}_\Delta(t^n)$ and its dissipation $\mathcal{I}_\Delta(t^n)$ obtained with the Scharfetter-Gummel extended scheme (**SGext**), the classical upwind scheme (**CU**) and our first and second order schemes (**FU1**)-(**FU2**). The classical upwind flux (**CU**) does not preserve the thermal equilibrium, which explains the phenomenon of saturation observed with it. The Scharfetter-Gummel extended flux (**SGext**) preserves the equilibrium at the points where the densities N and P do not vanish, but due to the zero boundary conditions on the left for N and on the right for P, there is also a phenomenon of saturation with it. Contrary to these two schemes, our new schemes (**FU1**)-(**FU2**) which preserve the equilibrium everywhere, provide a satisfying long-time behavior. Moreover, we computed the relative energy and its dissipation with our schemes for different numbers N_x of cells and notice that the decay rate does not depend on the spatial step size. We obtained satisfying results even for few number of cells.

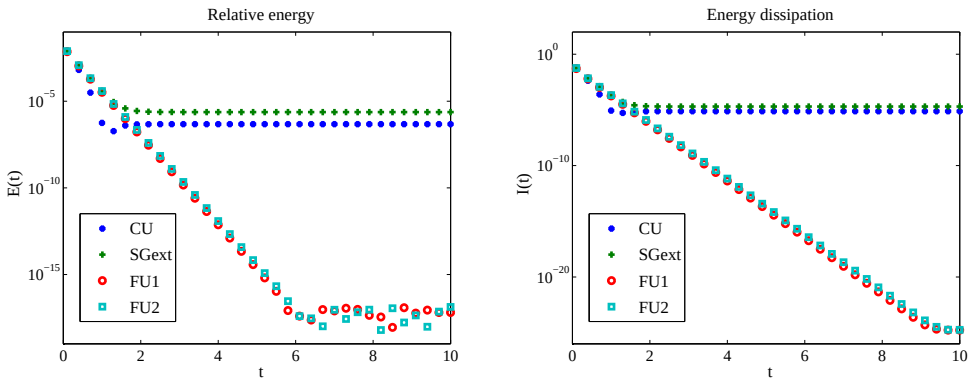


Figure 4.4: Example 4 - Evolution of the relative energy $\mathcal{E}_\Delta(t^n)$ and its dissipation $\mathcal{I}_\Delta(t^n)$ in log-scale for different schemes ($N_x = 64$).

Example 5. Let us consider now a 2D test case picked on the paper of C. Chainais-Hillairet, J. G. Liu and Y. J. Peng [52]. As in the previous example, the Dirichlet boundary conditions vanish on some part of the boundary. The time step is $\Delta t = 10^{-4}$, the final time is $T = 10$ and we compute an approximate solution on a 32×32 Cartesian grid.
In Figure 4.5, we compare the discrete relative energy $\mathcal{E}_\Delta(t^n)$ and its dissipation $\mathcal{I}_\Delta(t^n)$ obtained with the Scharfetter-Gummel extended scheme (**SGext**), the classical upwind scheme (**CU**) and the fully upwind schemes (**FU1**)-(**FU2**). We make the same observations as in Example 4: there is a phenomenon of saturation with the Scharfetter-Gummel extended and the classical upwind schemes, and not with our new scheme. Moreover, the decay rate does not depend on the number of grid cells chosen.

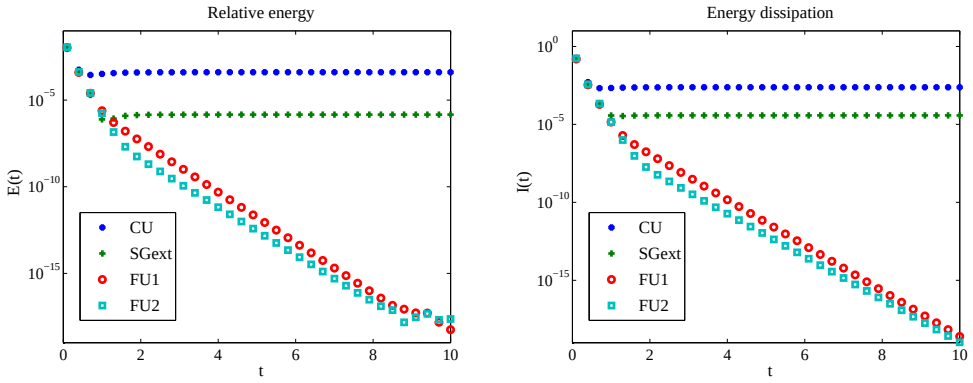


Figure 4.5: Example 5 - Evolution of the relative energy $\mathcal{E}_\Delta(t^n)$ and its dissipation $\mathcal{I}_\Delta(t^n)$ in log-scale for different schemes.

4.4.3 The porous media equation

In this subsection we approximate solutions to the porous media equation

$$\partial_t u = \nabla \cdot (xu + \nabla u^m).$$

We define an approximation $(U_i^{eq})_{i=1,...,N_x}$ of the unique stationary solution u^{eq} (4.1.6) by

$$U_i^{eq} = \left(\overline{C} - \frac{m-1}{2m}\,|x_i|^2\right)_+^{1/(m-1)}, \quad i = 1, ..., N_x,$$

where $\overline{C}$ is such that the discrete mass of $(U_i^{eq})_{i=1,...,N_x}$ is equal to that of $(U_i^0)_{i=1,...,N_x}$, namely
$\sum_i \Delta x_i U_i^{eq} = \sum_i \Delta x_i U_i^0$. We use a fixed point algorithm to compute this constant $\overline{C}$.

Example 6. We consider the following one dimensional test case: $m = 5$, with initial condition

$$u_0(x) = \begin{cases} 1 & \text{if } x \in (-3.7, -0.7) \cup (0.7, 3.7), \\ 0 & \text{otherwise.} \end{cases}$$

Then we compute the approximate solution on $(-5.5, 5.5)$, which is divided into $N_x = 160$ uniform cells. The time step is fixed to $\Delta t = 10^{-4}$ and the final time is $T = 10$.
In Figure 4.6 we compare the discrete relative entropy $\mathcal{E}_\Delta(t^n)$ and its dissipation $\mathcal{I}_\Delta(t^n)$ obtained with the Scharfetter-Gummel extended scheme, the classical upwind scheme and the first and second order fully upwind schemes. We obtain almost the same behavior for the Scharfetter-Gummel scheme and the fully upwind schemes. We only notice that the dissipation $\mathcal{I}_\Delta(t^n)$ obtained with the Scharfetter-Gummel scheme saturates before those obtained with the fully upwind schemes. This phenomenon of saturation is still greater for the classical upwind scheme. Moreover, we observe an exponential decay of $\mathcal{E}_\Delta(t^n)$ and $\mathcal{I}_\Delta(t^n)$, at a rate -12. In their paper [48], J. A. Carrillo and G. Toscani obtain the following equation for the entropy dissipation:

$$\frac{d}{dt}\mathcal{I}(t) = -2\,\mathcal{I}(t) - \mathcal{R}(t),$$

where $\mathcal{R}(t) \geq 0$ depends on the power m. Then they conclude with the exponential decay of the relative entropy $\mathcal{E}$ to zero at a rate -2. In our test where the initial condition is symmetric, we obtain a better rate, which seems to depend on m, taken equal to 5 here, thus it underlines the contribution of the term $\mathcal{R}$ in this case.
However, if we now consider a nonsymmetric initial data $u_0(x) = \mathbf{1}_{[2,3]}(x)$ and compute the relative entropy $\mathcal{E}_\Delta(t^n)$ obtained with our scheme (**FU2**) for different values of m, we observe in Figure 4.7 an exponential decay with rate -2, independently of the value of m. Thus in this case the estimate of decay of the relative entropy seems sharp.

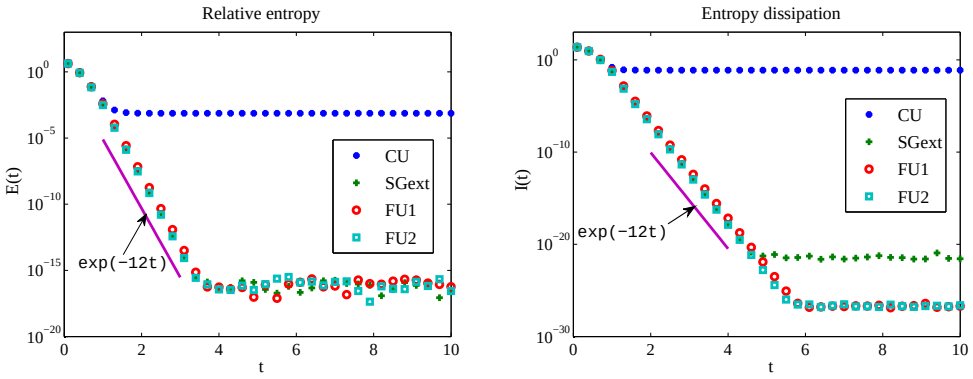


Figure 4.6: Example 6 - Evolution of the relative entropy $\mathcal{E}_\Delta(t^n)$ and its dissipation $\mathcal{I}_\Delta(t^n)$ in log-scale for different schemes.

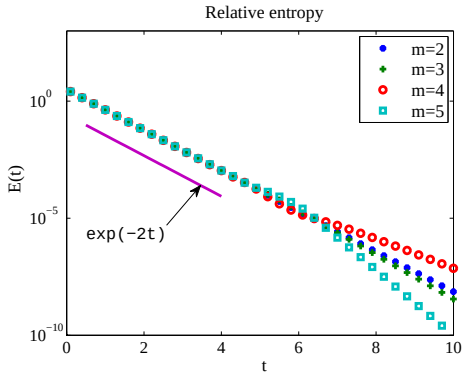


Figure 4.7: Evolution of the relative entropy $\mathcal{E}_\Delta(t^n)$ in log-scale for different values of m in the case of a nonsymmetric initial data.

Example 7. We still consider the porous media equation, but now in two space dimension on $\Omega = (-10, 10) \times (-10, 10)$. We take $m = 4$ and the initial condition is

$$u_0(x,y) = \begin{cases} \exp\left(-\frac{1}{6-(x-2)^2-(y+2)^2}\right) & \text{if } (x-2)^2+(y+2)^2 < 6, \\ \exp\left(-\frac{1}{6-(x+2)^2-(y-2)^2}\right) & \text{if } (x+2)^2+(y-2)^2 < 6, \\ 0 & \text{otherwise.} \end{cases}$$

We compute the approximate solution on a 200×200 Cartesian grid, with $\Delta t = 10^{-4}$ and $T = 10$.
In Figure 4.8 we compare the discrete relative entropy $\mathcal{E}_\Delta(t^n)$ and its dissipation $\mathcal{I}_\Delta(t^n)$ obtained with the Scharfetter-Gummel scheme, the classical upwind scheme and the fully upwind schemes, and obtain an exponential decay at a rate -4 with our new scheme (**FU2**). Figure 4.9 presents the evolution of the density of gas u computed with our second-order

scheme at four different times $t = 0$, $t = 0.5$, $t = 1$ and $t = 10$ and the approximation of the stationary solution u^{eq} corresponding to this initial data.

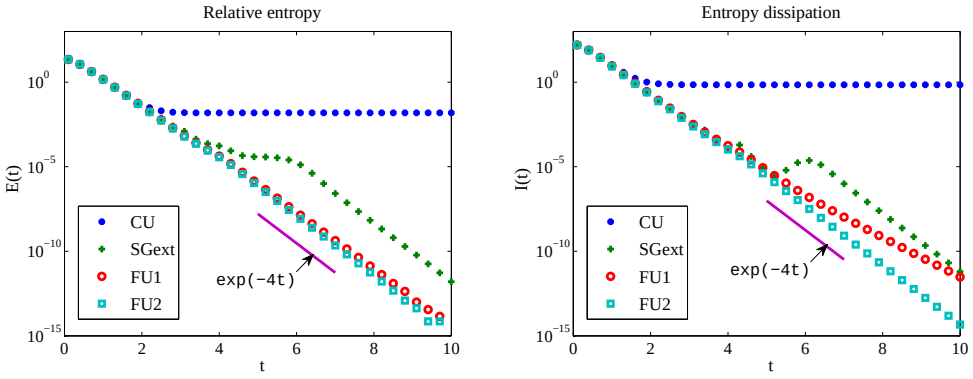


Figure 4.8: Example 7 - Evolution of the relative entropy $\mathcal{E}_\Delta(t^n)$ and its dissipation $\mathcal{I}_\Delta(t^n)$ in log-scale for different schemes.

4.4.4 Nonlinear Fokker-Planck equations for fermions and bosons

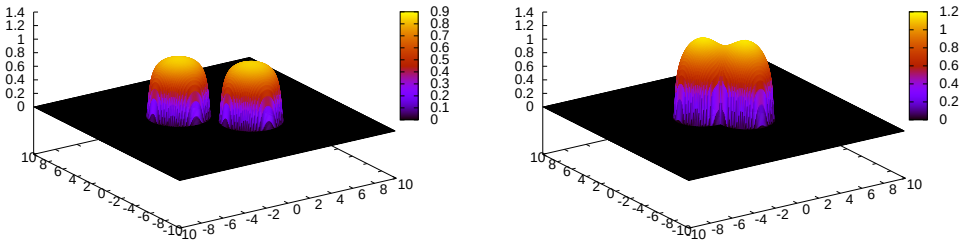
Example 8. We first consider the nonlinear Fokker-Planck equation (4.1.11) for fermions ($k = -1$). As in the porous media equation case, we define an approximation $(U_i^{eq})_{i=1,\dots,N_x}$ of the unique stationary solution u^{eq} (4.1.12) by

$$U_i^{eq} = \frac{1}{\overline{\beta} e^{\frac{|x_i|^2}{2}} + 1}, \quad i = 1, \dots, N_x,$$

where $\overline{\beta} \geq 0$ is such that the discrete mass of $(U_i^{eq})_{i=1,\dots,N_x}$ is equal to that of $(U_i^0)_{i=1,\dots,N_x}$. We use a fixed point algorithm to compute this constant $\overline{\beta}$.
We consider a 3D test case. The initial condition is chosen as the sum of four Gaussian distributions:

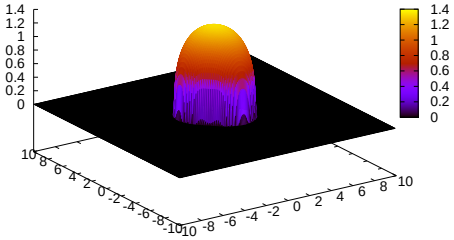
$$u_0(x) = \frac{1}{2\sqrt{2\pi}} \left(e^{-\frac{1}{2}|x-x_1|^2} + e^{-\frac{1}{2}|x-x_2|^2} + e^{-\frac{1}{2}|x-x_3|^2} + e^{-\frac{1}{2}|x-x_4|^2} \right),$$

where $x_1 = (2, 2, 2)$, $x_2 = (-2, -2, -2)$, $x_3 = (2, -2, 2)$ and $x_4 = (-2, 2, -2)$.
We consider a $40 \times 40 \times 40$ Cartesian grid of $\Omega = (-8, 8)^3$, $\Delta t = 10^{-4}$ and $T = 10$.
Evolution of the discrete relative entropy $\mathcal{E}_\Delta(t^n)$, its dissipation $\mathcal{I}_\Delta(t^n)$ and $\|U^n - U^{eq}\|_{L^1}$ obtained with the scheme (**FU2**) is presented in Figure 4.10. We observe exponential decay rate of these quantities, which is in agreement with the result proved by J. A. Carrillo, P. Laurençot and J. Rosado in [46]. In Figure 4.11 we report the evolution of the level set of the distribution function $u(t, x, y, z) = 0.1$ at different times and the level set of the corresponding equilibrium solution $u^{eq}(x, y, z) = 0.1$.

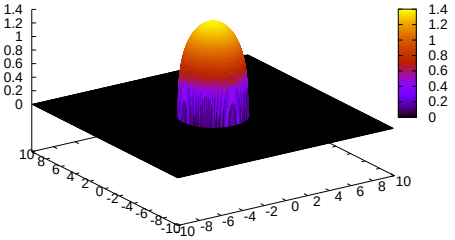


(a) $t = 0$

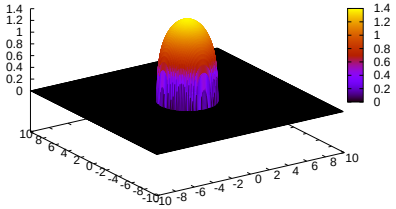
(b) $t = 0.5$



(c) $t = 1$



(d) $t = 10$



(e) Stationary solution

Figure 4.9: Example 7 - Evolution of the density of gas u and corresponding stationary solution u^{eq}.

Example 9. We now consider the more general Fokker-Planck equation (4.1.13) with $N = 3$ in 1D:

$$\partial_t u = \partial_x(xu(1 + u^3) + \partial_x u).$$

The initial condition is given by the sum of two Gaussian distributions:

$$u_0(x) = \frac{M}{2\sqrt{2\pi}} \left(\exp\left(-\frac{|x-2|^2}{2} \right) + \exp\left(-\frac{|x+2|^2}{2} \right) \right),$$

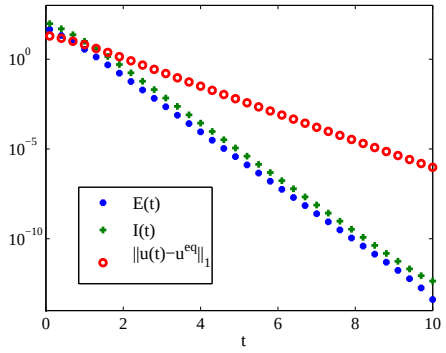


Figure 4.10: Example 8 - Evolution of the relative entropy $\mathcal{E}_\Delta(t^n)$, the dissipation $\mathcal{I}_\Delta(t^n)$ and the L^1 norm $\|U^n - U^{eq}\|_1$.

where $M \geq 0$ is the mass of u_0. We compute an approximate solution with the scheme (**FU2**) for two different values of M. The computational domain $(-10, 10)$ is divided into $N_x = 500$ uniform cells.
According to the paper of N. Ben Abdallah, I. Gamba and G. Toscani [15], there is a phenomenon of critical mass in this case. In Figure 4.12, we represent the evolution of the density u until time $T = 10$ for an initial sub-critical mass $M = 1$. We observe the convergence of the solution to the unique minimizer u^{eq} of the entropy functional, given by

$$u^{eq}(x) = \left(\beta\, e^{3x^2/2} - 1\right)^{-\frac{1}{3}},$$

according to [15], where β is such that $\int u^{eq}(x)\, dx = M$. Moreover, we observe in this case an exponential decay rate of the dissipation and the L^1 distance between the solution and the equilibrium.
In Figure 4.13, we represent the evolution of the density u for an initial super-critical mass $M = 10$ until time $T = 0.9$. We observe in this case the convergence of the solution to an equilibrium which has a singular part localized in the origin, which is in agreement with the result proved in [15].

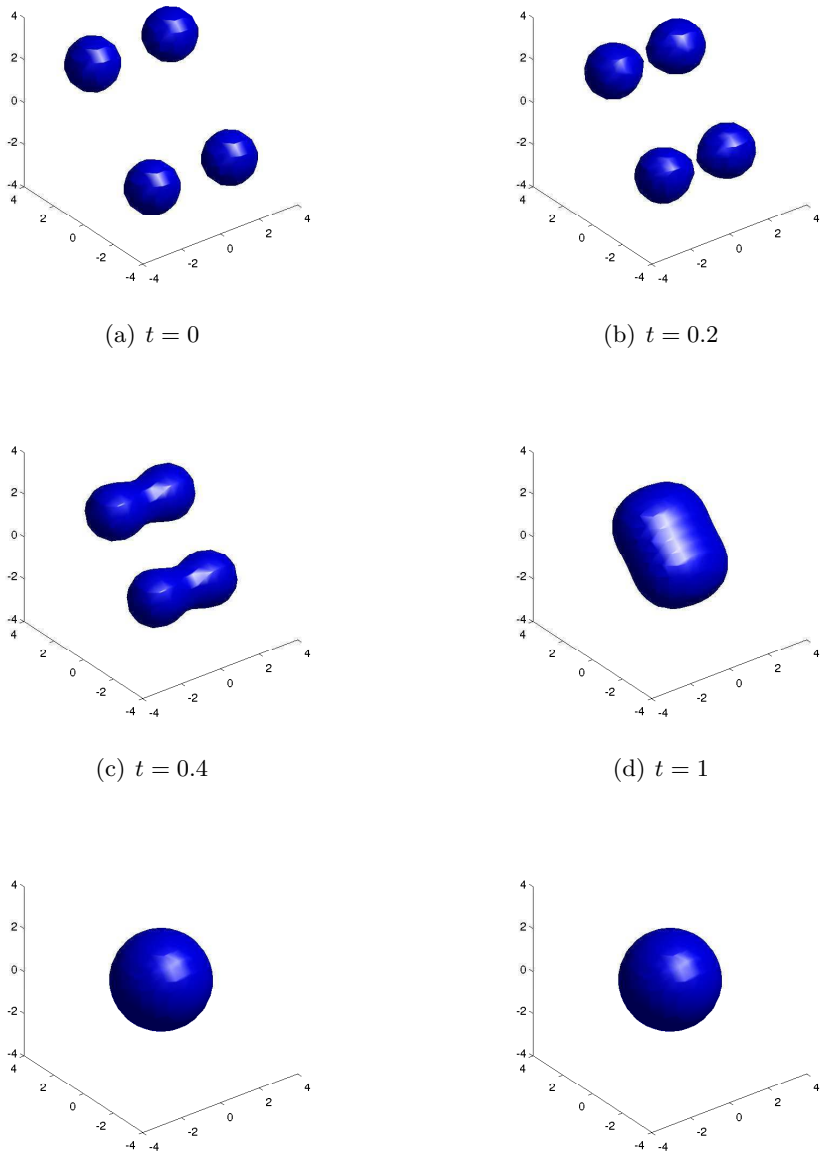
4.4.5 The Buckley-Leverett equation

Finally we consider the Buckley-Leverett equation, with both nonlinear convection and diffusion:

$$\partial_t u = \partial_x \left(-f(u) + \partial_x r(u)\right). \tag{4.4.1}$$

The Buckley-Leverett equation is a simple model for displacement of oil by water in oil reservoirs. The function $u(t, x) \in [0, 1]$ represents the fraction of fluid corresponding to oil. We consider a fractional flow function f with a s-shaped form

$$f(u) = \frac{u^2}{u^2 + (1-u)^2}.$$



(a) $t = 0$ (b) $t = 0.2$

(c) $t = 0.4$ (d) $t = 1$

(e) $t = 10$ (f) Stationary solution

Figure 4.11: Example 8 - Evolution of the level set $u(t, x, y, z) = 0.1$ and level set of the corresponding stationary solution $u^{eq}(x, y, z) = 0.1$.

This choice corresponds to a model which does not include the gravitational effects. The function r is such that $r'(u) = \varepsilon\, \nu(u)$, where the capillary diffusion coefficient is given by

$$\nu(u) = 4u(1 - u).$$

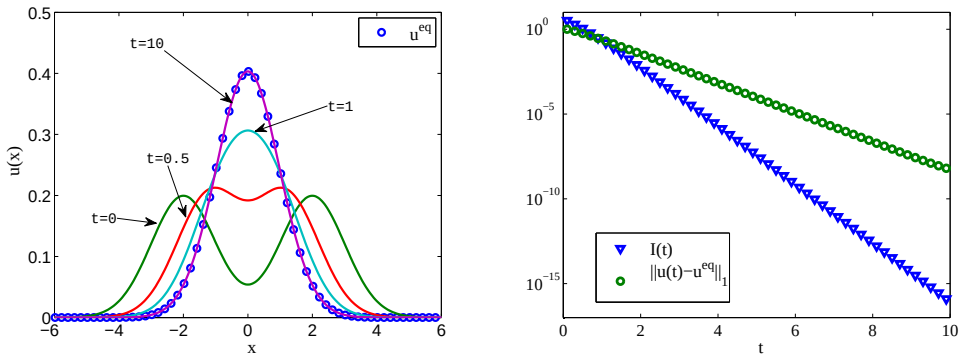
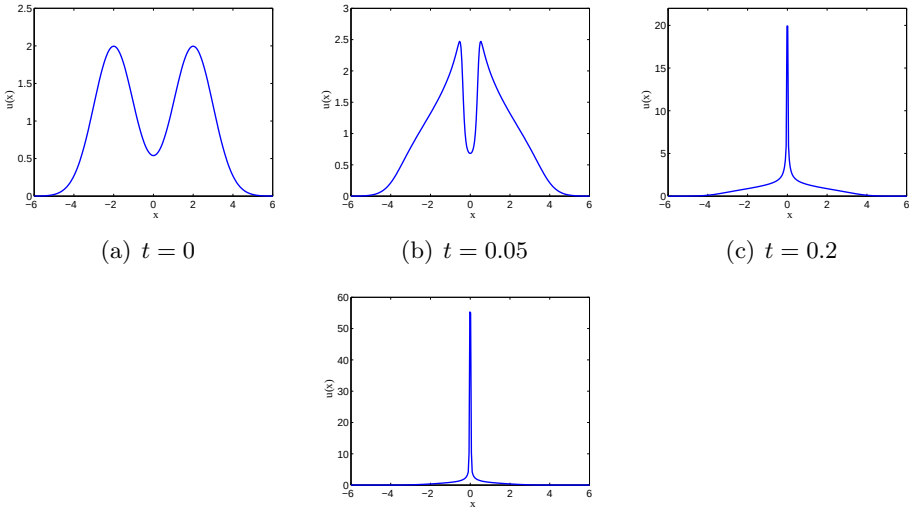


Figure 4.12: Example 9 - Evolution of the density u of sub-critical mass $M = 1$ (left) and of the corresponding dissipation $\mathcal{I}_\Delta(t^n)$ and L^1 norm $\|U^n - U^{eq}\|_1$ (right).



(a) $t = 0$ (b) $t = 0.05$ (c) $t = 0.2$

(d) $t = 0.9$

Figure 4.13: Example 9 - Evolution of the density u of super-critical mass $M = 10$.

The scaling parameter $\varepsilon > 0$ in front of the capillary diffusion is usually small.
In this particular case, the Buckley-Leverett equation (4.4.1) possesses a functional which dissipates a quantity. Indeed, rewriting the flux under the form (4.2.4) by taking $V = -x$ and

$$h(u) = 4\left(\log(u) - 3u + 2u^2 - \frac{2}{3}u^3\right),$$

multiplying the equation (4.4.1) by $(-x + h(u))$ and integrating over Ω, we get

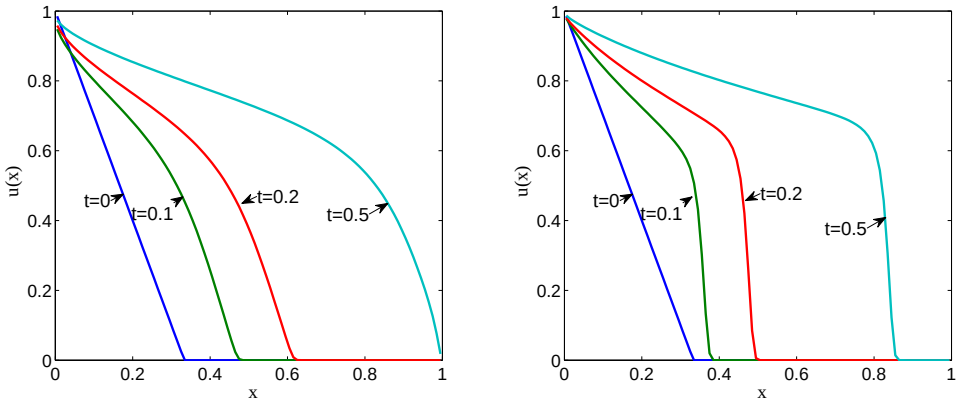
$$\frac{d}{dt}\int_\Omega (-x + h(u))\, u\, dx = -\int_\Omega f(u)\, |\partial_x (-x + h(u))|^2\, dx \leq 0,$$

since $f(u) \geq 0$ for all $u \in [0, 1]$.

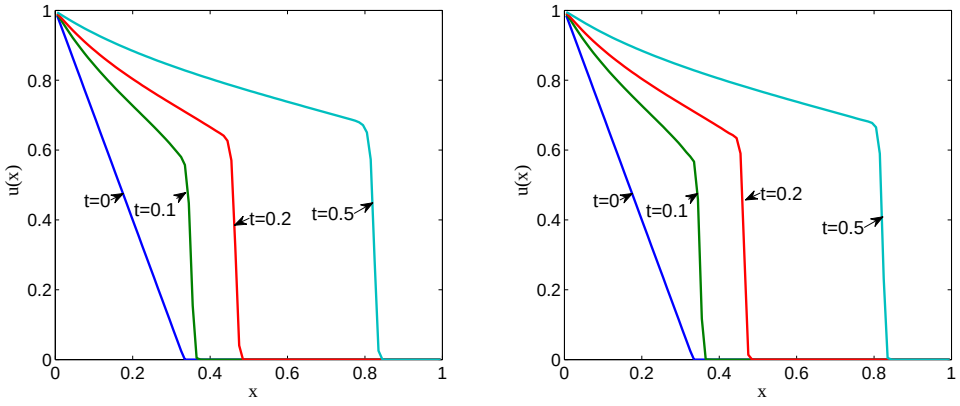
Example 10. We consider the following test case [130, 135]: the domain Ω is $(0,1)$, the initial condition

$$u_0(x) = \begin{cases} 1-3x & \text{if} \quad 0 \le x \le \frac{1}{3}, \\ \\ 0 & \text{if} \quad \frac{1}{3} < x \le 1, \end{cases}$$

and the boundary condition $u(0,t) = 1$. The domain is divided into $N_x = 100$ cells and the time step is $\Delta t = 10^{-4}$. The numerical solution computed at different times for different values of ε is shown in Figure 4.14. The results compare well with those in [130, 135]. Moreover, our scheme remains valid for all values of ε, even $\varepsilon = 0$. In this case the fully upwind flux degenerates into the well-known local Lax-Friedrichs flux.



(a) $\varepsilon = 10^{-1}$ (b) $\varepsilon = 10^{-2}$



(c) $\varepsilon = 10^{-3}$ (d) $\varepsilon = 0$

Figure 4.14: Example 10 - Evolution of the numerical solution for different values of ε.

4.5 Conclusion

In this chapter we have presented how to build a new finite volume scheme for nonlinear degenerate parabolic equations which admit an entropy functional. To this end, we rewrite the equation in the form of a convection equation, by taking the convective and diffusive parts into account together. Then we apply either the upwind method in the linear case or the local Lax-Friedrichs method in the nonlinear case.
On the one hand, this construction ensures that a particular type of steady-state is preserved. We obtain directly a semi-discrete entropy estimate, which is the first step to prove the large-time behavior of the numerical solution. On the other hand, we use a slope-limiter method to get second-order accuracy even in the degenerate case.
Numerical examples demonstrate high-order accuracy of the scheme. Moreover we have applied it to some of the physical models for which the long-time behavior has been studied: the porous media equation, the drift-diffusion system for semiconductors, the nonlinear Fokker-Planck equation for bosons and fermions. We obtain the convergence of the approximate solution to an approximation of the equilibrium state at an exponential rate. A future work would be to prove this exponential rate by using a discrete entropy/entropy dissipation estimate as in the continuous case compared with previous approaches.

Troisième partie

Un schéma volumes finis pour un modèle de chimiotactisme avec diffusion croisée

Dans cette dernière partie, nous analysons un schéma volumes finis pour un modèle de Keller-Segel en dimension 2 avec diffusion croisée étudié dans [111] :

$$\text{(4.5.1)} \qquad \begin{cases} \partial_t n &= \operatorname{div}(\nabla n - n\nabla S), \\ \alpha\,\partial_t S &= \Delta S + \delta\Delta n + \mu\, n - S, \end{cases}$$

où $n(x,t)$ désigne la densité de cellules et $S(x,t)$ la concentration en chimioattractant. L'ajout dans l'équation sur S du terme de diffusion croisée $\delta\Delta n$, $\delta > 0$, empêche l'explosion des solutions en temps fini qui peut avoir lieu pour le système de Keller-Segel classique (voir par exemple [112, 116]). Il est prouvé dans [111] en utilisant une fonctionnelle d'entropie que le système avec diffusion croisée (4.5.1) admet une solution faible globale pour des valeurs de $\delta > 0$ arbitrairement petites. De plus dans le cas parabolique-elliptique ($\alpha = 0$), si μ est suffisamment grand ou si δ est suffisamment petit, la convergence en temps long à vitesse exponentielle vers l'état stationnaire homogène est démontrée.

Nous proposons ici un schéma volumes finis pour le système (4.5.1) dans le cas $\alpha = 0$ (parabolique-elliptique) dont l'analyse repose sur une adaptation au cadre discret des techniques utilisées dans le contexte continu. L'étude du schéma numérique s'appuie notamment sur une estimation d'entropie discrète, dont l'établissement nécessite l'utilisation de versions discrètes d'inégalités fonctionnelles telles que

– l'inégalité de Gagliardo-Nirenberg-Sobolev : Ω étant un domaine borné de $\mathbb{R}^N$, $N \geq 2$, pour $1 < p, q \leq \infty$, il existe une constante $C > 0$ telle que pour tout $u \in W^{1,p}(\Omega) \cap L^q(\Omega)$,
$$\| u \|_{L^m(\Omega)} \leq C \| u \|^{\theta}_{W^{1,p}(\Omega} \| u \|^{1-\theta}_{L^q(\Omega)},$$
où
$$0 \leq \theta \leq \frac{p}{p + q(p-1)} \leq 1 \quad \text{et} \quad \frac{1}{m} = \frac{1-\theta}{q} + \frac{\theta}{p} - \frac{\theta}{N},$$

– l'inégalité de Sobolev-Poincaré :
$$\| u \|_{L^q(\Omega)} \leq C \| u \|_{W^{1,p}(\Omega)} \quad \forall u \in W^{1,p}(\Omega),$$
pour
$$1 \leq q \leq \frac{pN}{N-p} \quad \text{si} \quad 1 \leq p < N,$$
$$\text{ou} \quad 1 \leq q < +\infty \qquad \text{si} \quad p \geq N.$$

Ces inégalités sont un outil standard pour l'étude de l'existence et de la régularité de solutions d'équations aux dérivées partielles elliptiques et paraboliques. Le cadre L^2 est généralement utilisé pour les problèmes elliptiques linéaires, tandis que le cadre L^p est crucial pour l'étude d'équations elliptiques ou paraboliques non linéaires, pour obtenir des estimations d'énergie permettant de montrer l'existence de solutions faibles. Dans le cadre discret, on a souvent besoin d'analogues de ces inégalités pour analyser la convergence des méthodes numériques et obtenir des estimations d'erreur.

Des versions discrètes de ces inégalités, classiques dans le cadre continu, sont démontrées dans plusieurs travaux, mais dans des contextes particuliers ou avec des preuves assez techniques (voir par exemple [85, 97]). En se fondant sur une idée présentée dans [89], à savoir l'injection continue de $BV(\Omega)$ dans $L^{N/(N-1)}(\Omega)$ pour un domaine Lipschitz $\Omega \subset \mathbb{R}^N$, nous proposons dans le chapitre 5 une étude dans un cadre plus général de ces inégalités, incluant notamment le cas de conditions aux limites mixtes. Nous présentons également une extension de ces résultats au contexte des schémas DDFV ("Discrete Duality Finite Volume"), bien adaptés pour l'approximation de problèmes elliptiques anisotropes sur des maillages généraux en 2D et 3D (voir par exemple [5, 70, 73]).

Dans le chapitre 6, nous nous intéressons à l'étude proprement dite du schéma volumes finis pour le modèle (4.5.1), dans le cas $\alpha = 0$. Le schéma numérique proposé est le même que dans [89], avec un terme de diffusion croisée additionnel. Nous étudions la convergence du schéma vers la solution faible du système. Cette analyse est fondée sur des estimations a priori obtenues grâce à l'inégalité d'entropie discrète, qui permettent d'établir la compacité d'une famille de solutions approchées. De plus, dans le cas d'un paramètre de diffusion croisée δ suffisamment grand, nous prouvons la convergence en temps long de la solution approchée vers l'état stationnaire homogène. Enfin, des simulations numériques pour des valeurs intermédiaires du paramètre δ semblent indiquer l'existence de solutions stationnaires non homogènes.

Chapitre 5

Quelques inégalités fonctionnelles discrètes pour des schémas volumes finis *

Nous prouvons des inégalités de type Gagliardo-Nirenberg-Sobolev et Poincaré-Sobolev pour des approximations obtenues sur des maillages de type volumes finis, avec des conditions au bord arbitraires. Le point-clé de la démonstration de ces inégalités est l'utilisation de l'injection continue de l'espace des fonctions à variation bornée $BV(\Omega)$ dans $L^{N/(N-1)}(\Omega)$, pour Ω un domaine Lipschitzien de $\mathbb{R}^N$, où $N \geq 2$. Finalement, nous appliquons également ces techniques dans le contexte des schémas DDFV ("Discrete Duality Finite Volume"), qui sont utilisés pour l'approximation de problèmes elliptiques et paraboliques non linéaires et anisotropes.

∗. Ce chapitre est un travail réalisé en collaboration avec C. Chainais-Hillairet et F. Filbet, *On discrete functional inequalities for some finite volume schemes* [17], soumis pour publication.

Contents

5.1 Introduction

In this chapter, we establish some discrete functional inequalities which are sometimes useful for the convergence analysis of finite volume schemes. In the continuous framework, the Gagliardo-Nirenberg-Sobolev and Sobolev-Poincaré inequalities are fundamental for the analysis of partial differential equations. They are a standard tool in existence and regularity theories for solutions. The L^2 framework is generally used for linear elliptic problems, more precisely it is a classical way to prove the coercivity of bilinear forms in H_0^1, which then allows to apply the Lax-Milgram theorem to prove existence of weak solutions. More generally, the L^p framework is crucial for the study of nonlinear elliptic or parabolic equations, to obtain some energy estimates which are useful to prove existence of weak solutions. Poincaré-type inequalities are also one of the step in the study of convergence to equilibrium for kinetic equations.

5.1.1 Gagliardo-Nirenberg-Sobolev and Sobolev-Poincaré inequalities

In the continuous situation, the Gagliardo-Nirenberg-Sobolev inequality writes as follows. Let assume $N \geq 2$ and Ω be an open bounded domain of $\mathbb{R}^N$. Then for $1 \leq p, q \leq \infty$, there exists a constant $C > 0$ such that for all $u \in W^{1,p}(\Omega) \cap L^q(\Omega)$,

$$\|u\|_{L^m(\Omega)} \leq C \|u\|^{\theta}_{W^{1,p}(\Omega)} \|u\|^{1-\theta}_{L^q(\Omega)}, \tag{5.1.1}$$

where

$$0 \leq \theta \leq 1 \quad \text{and} \quad \frac{1}{m} = \frac{1-\theta}{q} + \frac{\theta}{p} - \frac{\theta}{N}.$$

We refer to [90, 144] for a proof of this result. We also remind the well-known Sobolev-Poincaré inequality [1, 31]:

$$\|u\|_{L^q(\Omega)} \leq C \|u\|_{W^{1,p}(\Omega)} \quad \forall u \in W^{1,p}(\Omega), \tag{5.1.2}$$

for

$$1 \leq q \leq \frac{pN}{N-p} \quad \text{if} \quad 1 \leq p < N,$$

$$\text{or} \quad 1 \leq q < +\infty \qquad \text{if} \quad p \geq N.$$

The mathematical analysis of convergence and error estimates for numerical methods are performed using functional analysis tools, such as discrete Sobolev inequalities. Several Poincaré-Sobolev inequalities have been established for the finite volume schemes as well as for the finite element methods. Concerning the finite volume framework, the first estimates were obtained in the particular case $N = 2$, $p = q = 2$ (which is the standard Poincaré inequality) for Dirichlet boundary conditions by Y. Coudière, J.-P. Vila and P. Villedieu [63]. The idea of the proof in these papers is to use some geometrical properties of the mesh. More precisely, given an oriented direction $\mathcal{D}$, any cell center of the mesh is connected to an upstream (with respect to $\mathcal{D}$) center of an edge of the boundary $\partial\Omega$ by a straight line of direction $\mathcal{D}$. This connection crosses a certain number of cells and their interfaces, and this argument allows to link a norm of the piecewise constant function considered with a norm of a discrete version of its gradient. This result was later generalized to the case of dimension $N = 3$ by R. Eymard, T. Gallouët and R. Herbin [85]. Also the same method has been applied to get more general Sobolev-Poincaré inequalities (5.1.2) for $1 \leq p = q \leq 2$ by J. Droniou, T. Gallouët and R. Herbin [75], and for $p = 2$, $1 \leq q < \infty$ if $N = 2$, $1 \leq q \leq 6$ if $N = 3$ by Y. Coudière, T. Gallouët and R. Herbin [62], still in the case of Dirichlet boundary conditions. Concerning the case of Neumann boundary conditions, a discrete Poincaré-Wirtinger inequality ($p = q = 2$) was established in [85, 93] for $N = 2$ or 3 by using the same method, as well as more general Sobolev inequalities ($p = 2$, $1 \leq q < \infty$ if $N = 2$, $1 \leq q \leq 6$ if $N = 3$) in [50].

More recently, another idea was used to prove this type of discrete inequalities: the continuous embedding of $BV(\Omega)$ into $L^{N/(N-1)}(\Omega)$ for a Lipschitz domain Ω. This argument was first exploited in [89] to prove a discrete Sobolev-Poincaré inequality (5.1.2) in dimension $N = 2$ with $q = 2$ and $p = 1$, in the case of Neumann boundary conditions. Then this method was used in [86] to prove general Sobolev-Poincaré inequalities (5.1.2) in any dimension $N \geq 1$ in the particular case of homogeneous Dirichlet boundary conditions. This result was then adapted to the case of Neumann or mixed boundary conditions by B. Andreianov, M. Bendahmane and R. Ruiz Baier in [4]. We also mention [25] where the continuous embedding of $BV(\mathbb{R}^N)$ into $L^{N/(N-1)}(\mathbb{R}^N)$ is used to establish an improved discrete Gagliardo-Nirenberg-Sobolev inequality in the whole space $\mathbb{R}^N$, $N \geq 1$.

Finally for $p = 2$, general discrete Sobolev-Poincaré inequalities are obtained in [97] for Voronoi finite volume approximations in the case of arbitrary boundary conditions by

using an adaptation of Sobolev's integral representation and the Voronoi property of the mesh. Concerning the finite element framework, a variant of a Poincaré-type inequality ($p = q = 2$) for functions in broken Sobolev spaces was derived in [12] for $N = 2$ and in [30, 163] for $N = 2, 3$. Then a generalised result was proposed in [131], providing bounds on the L^q norms in terms of a broken H^1 norm ($p = 2$, $1 \leq q < \infty$ if $N = 2$ and $1 \leq q \leq 2N/(N-2)$ if $N \geq 3$). The proof is based on elliptic regularity results and nonconforming finite element interpolants. Finally, a result in non-Hilbertian setting ($p \neq 2$) was obtained in [72], taking inspiration from the technique used by F. Filbet [89] and also R. Eymard, T. Gallouët and R. Herbin [86], namely the continuous embedding of $BV(\Omega)$ into $L^{N/(N-1)}(\Omega)$.

5.1.2 Aim of the chapter and outline

In this chapter our aim is to provide a simple proof to discrete versions of Gagliardo-Nirenberg-Sobolev (5.1.1) and Sobolev-Poincaré (5.1.2) inequalities for functions coming from finite volume schemes with arbitrary boundary values. Several Sobolev-Poincaré inequalities are already proved as mentioned above but here we propose a unified result. It includes in particular the case of mixed boundary conditions. Concerning Gagliardo-Nirenberg-Sobolev inequalities, the result of F. Bouchut, R. Eymard and A. Prignet [25] is to our knowledge the only available, and it deals with the case of the whole space $\mathbb{R}^N$.

Our starting point to prove these discrete estimates is the continuous embedding of $BV(\Omega)$ into $L^{N/(N-1)}(\Omega)$, as in [89, 86, 72, 25]. The main difficulty appears when boundary conditions must be taken into account. In the papers mentioned previously [89, 86, 72], the boundary conditions are either homogeneous Dirichlet or Neumann on the whole boundary. In [25], the problem is considered in the whole space $\mathbb{R}^N$. In the case where the function satisfies homogeneous Dirichlet boundary conditions only on a part $\Gamma^0 \subsetneq \partial\Omega$ of the boundary, we cannot use the same strategy as in [86], which consists of extending the function considered to $\mathbb{R}^N$ by zero. Our idea is to thicken the boundary of Ω to take the mixed boundary conditions into account in this case.

The outline of the chapter is as follows. In Section 5.2, we first define the functional spaces: the space of finite volume approximations and the space $BV(\Omega)$. We will see that $BV(\Omega)$ is a natural space to study piecewise constant functions as finite volume approximations. In Section 5.3, we do not take into account any boundary conditions and prove the discrete Gagliardo-Nirenberg-Sobolev inequalities (Theorem 5.3.1) and the discrete Sobolev-Poincaré inequalities (Theorem 5.3.2) in this case. These results are the discrete counterpart of (5.1.1) and (5.1.2). They may be used for instance in the convergence analysis of finite volume schemes in the case with Neumann boundary conditions. Then, in Section 5.4, we consider the case where the discrete function is given by a finite volume scheme with homogeneous boundary conditions on a part of the boundary. In this case, the discrete space (for the finite volume approximations) is unchanged. However, the discrete $W^{1,p}$ seminorm will take into account some jumps on the boundary. We prove discrete Gagliardo-Nirenberg-Sobolev inequalities (Theorem 5.4.1) and discrete Sobolev-Poincaré inequalities (Theorem 5.4.2), similar to (5.1.1) and (5.1.2) but with the $W^{1,p}$-seminorm

instead of the full $W^{1,p}$-norm. Finally, in Section 5.5, we show how to extend the results from Sections 5.3 and 5.4 to finite volume approximations coming from discrete duality finite volume (DDFV) schemes. This family of schemes is mainly applied to elliptic and parabolic problems. This method can be applied to a wide class of 2D meshes (but also 3D [61]) and inherits the main qualitative properties of the continuous problem: monotonicity, coercivity, variational formulation, etc...

5.2 Functional spaces

5.2.1 The space of finite volume approximations

We now introduce the discrete settings, including notations and assumptions on the meshes and definitions of the discrete norms. Let Ω be an open bounded polyhedral susbset (Lipschitz domain) of $\mathbb{R}^N$, $N \geq 2$, and $\Gamma := \partial\Omega$ its boundary. An admissible mesh of Ω is given by a family $\mathfrak{M}$ of control volumes, a family $\mathcal{E}$ of relatively open parts of hyperplans in $\mathbb{R}^N$ (which represent the faces of the control volumes) and a family of points $(x_K)_{K\in\mathfrak{M}}$ which satisfy the following properties:

- $\overline{\Omega} = \bigcup_{K\in\mathfrak{M}} \overline{K}$,
- for all $K \in \mathfrak{M}$, there exists $\mathcal{E}_K \subset \mathcal{E}$ such that $\partial K = \bigcup_{\sigma\in\mathcal{E}_K} \overline{\sigma}$,
- for all $(K, L) \in \mathfrak{M}^2$ with $K \neq L$, either $\mathrm{m}(\overline{K} \cap \overline{L}) = 0$, or $\overline{K} \cap \overline{L} = \overline{\sigma}$ for some $\sigma \in \mathcal{E}$, which will be denoted by $K|L$,
- the family of points $(x_K)_{K\in\mathcal{T}}$ is such that for all $K \in \mathfrak{M}$, $x_K \in K$ and if $\sigma = K|L$, it is assumed that $x_K \neq x_L$.

In the set of faces $\mathcal{E}$, we distinguish the interior faces $\sigma \in \mathcal{E}_{int}$ and the boundary faces $\sigma \in \mathcal{E}_{ext}$. For a control volume $K \in \mathfrak{M}$, we denote by $\mathcal{E}_K$ the set of its faces, $\mathcal{E}_{int,K}$ the set of its interior faces and $\mathcal{E}_{ext,K}$ the set of faces of K included in the boundary Γ.
In the sequel we denote by d the distance in $\mathbb{R}^N$, m the Lebesgue measure in $\mathbb{R}^N$ or $\mathbb{R}^{N-1}$. For all $\sigma \in \mathcal{E}$, we define

$$d_\sigma = \begin{cases} \mathrm{d}(x_K, x_L) & \text{for} \quad \sigma = K|L \in \mathcal{E}_{int}, \\ \mathrm{d}(x_K, \sigma) & \text{for} \quad \sigma \in \mathcal{E}_{ext,K}. \end{cases}$$

We assume that the family of meshes considered satisfies the following regularity constraint: there exists $\xi > 0$ such that

$$\mathrm{d}(x_K, \sigma) \geq \xi\, d_\sigma, \quad \text{for } K \in \mathfrak{M}, \quad \text{for } \sigma \in \mathcal{E}_K. \tag{5.2.1}$$

The size of the mesh is defined by

$$h = \max_{K\in\mathfrak{M}} (\mathrm{diam}(K)). \tag{5.2.2}$$

In general, finite volume methods lead to the computation of one discrete unknown by control volume. The corresponding finite volume approximation is a piecewise constant

function. Therefore, we define the set $X(\mathfrak{M})$ of the finite volume approximation:

$$X(\mathfrak{M}) = \left\{ u \in L^1(\Omega) \ / \ \exists (u_K)_{K\in\mathfrak{M}} \text{ such that } u = \sum_{K\in\mathfrak{M}} u_K \mathbf{1}_K \right\}.$$

Let us now define some discrete norms and seminorms on $X(\mathfrak{M})$.

Definition 5.1. Let Ω be a bounded polyhedral subset of $\mathbb{R}^N$, $\mathfrak{M}$ an admissible mesh of Ω.

1. For $p \in [1, +\infty)$, the discrete L^p norm is defined by

$$\| u \|_{0,p,\mathfrak{M}} = \left(\sum_{K\in\mathfrak{M}} \mathrm{m}(K) \, | u_K |^p \right)^{\frac{1}{p}}, \quad \forall u \in X(\mathfrak{M}).$$

2. In the general case, for $p \in [1, +\infty)$, the discrete $W^{1,p}$-seminorm is defined by:

$$| u |_{1,p,\mathfrak{M}} = \left(\sum_{\substack{\sigma\in\mathcal{E}_{int} \\ \sigma = K|L}} \frac{\mathrm{m}(\sigma)}{d_\sigma^{p-1}} \, |u_L - u_K|^p \right)^{\frac{1}{p}}, \quad \forall u \in X(\mathfrak{M})$$

and the discrete $W^{1,p}$-norm is defined by

$$\| u \|_{1,p,\mathfrak{M}} = \| u \|_{0,p,\mathfrak{M}} + | u |_{1,p,\mathfrak{M}}, \quad \forall u \in X(\mathfrak{M}). \tag{5.2.3}$$

3. In the case where homogeneous Dirichlet boundary conditions are underlying (because the piecewise constant function comes from a finite volume scheme), we need to take into account jumps on the boundary in the discrete $W^{1,p}$-seminorm. Let $\Gamma^0 \subset \Gamma$ be a part of the boundary. In the set of exterior faces $\mathcal{E}_{ext}$, we distinguish $\mathcal{E}_{ext}^0$ the set of boundary faces included in Γ^0. For $p \in [1, +\infty)$, we define the discrete $W^{1,p}$-seminorm (which depends on Γ^0) by

$$| u |_{1,p,\Gamma^0,\mathfrak{M}} = \left(\sum_{\sigma\in\mathcal{E}} \frac{\mathrm{m}(\sigma)}{d_\sigma^{p-1}} \, (D_\sigma u)^p \right)^{\frac{1}{p}}, \quad 1 \le p < +\infty, \tag{5.2.4}$$

where

$$D_\sigma u = \begin{cases} | u_K - u_L | & \text{if} \quad \sigma = K|L \in \mathcal{E}_{int}, \\ | u_K | & \text{if} \quad \sigma \in \mathcal{E}_{ext}^0 \cap \mathcal{E}_K, \\ 0 & \text{if} \quad \sigma \in \mathcal{E}_{ext} \setminus \mathcal{E}_{ext}^0. \end{cases}$$

We then define the discrete $W^{1,p}$ norm by

$$\| u \|_{1,p,\mathfrak{M}} = \| u \|_{0,p,\mathfrak{M}} + | u |_{1,p,\Gamma^0,\mathfrak{M}}, \ \forall u \in X(\mathfrak{M}). \tag{5.2.5}$$

5.2.2 The space $BV(\Omega)$

Let us first recall some results concerning functions of bounded variation (we refer to [3, 167] for a thorough presentation $BV(\Omega)$). Let Ω be an open set of $\mathbb{R}^N$ and $u \in L^1(\Omega)$. The *total variation* of u in Ω, denoted by $TV_\Omega(u)$, is defined by

$$TV_\Omega(u) = \sup\left\{ \int_\Omega u(x) \operatorname{div}(\phi(x))\, dx, \quad \phi \in \mathcal{C}_c^1(\Omega), \quad |\phi(x)| \le 1, \quad \forall x \in \Omega \right\}$$

and the function $u \in L^1(\Omega)$ belongs to $BV(\Omega)$ if and only if $TV_\Omega(u) < +\infty$. The space $BV(\Omega)$ is endowed with the norm

$$\| u \|_{BV(\Omega)} := \| u \|_{L^1(\Omega)} + TV_\Omega(u).$$

The space $BV(\Omega)$ is a natural space to study finite volume approximations. Indeed, as it is proved for instance in [89], for $u \in X(\mathfrak{M})$, we have

$$TV_\Omega(u) \le \sum_{\substack{\sigma \in \mathcal{E}_{int} \\ \sigma = K|L}} \mathrm{m}(\sigma)\ |u_L - u_K| = |u|_{1,1,\mathfrak{M}} < +\infty.$$

The discrete space $X(\mathfrak{M})$ is included in $L^1 \cap BV(\Omega)$. Moreover, $\|u\|_{BV(\Omega)} \le \|u\|_{1,1,\mathfrak{M}}$.

Our starting point for the discrete functional inequalities is the continuous embedding of $BV(\Omega)$ into $L^{N/(N-1)}(\Omega)$ for a Lipschitz domain Ω, recalled in Theorem 5.2.1.

Theorem 5.2.1. *Let Ω be a Lipschitz bounded domain of $\mathbb{R}^N$, $N \ge 2$. Then there exists a constant $c(\Omega)$ only depending on Ω such that:*

$$\left(\int_\Omega | u |^{\frac{N}{N-1}}\, dx \right)^{\frac{N-1}{N}} \le c(\Omega)\ \| u \|_{BV(\Omega)}, \quad \forall u \in BV(\Omega). \tag{5.2.6}$$

There are also more precise results involving only the seminorm $TV_\Omega(u)$ instead of the norm $\| u \|_{BV(\Omega)}$. Indeed, the seminorm TV_Ω becomes a norm on the space of BV functions vanishing on a part of the boundary and also on the space of BV functions with a zero mean value. In these cases, the continuous embedding of $BV(\Omega)$ into $L^{N/(N-1)}(\Omega)$ rewrites as in Theorem 5.2.2.

Theorem 5.2.2. *Let Ω be a Lipschitz bounded connected domain of $\mathbb{R}^N$, $N \ge 2$.*

1. *There exists a constant $c(\Omega) > 0$ only depending on Ω such that, for all $u \in BV(\Omega)$,*

$$\left(\int_\Omega | u - \overline{u} |^{\frac{N}{N-1}}\, dx \right)^{\frac{N-1}{N}} \le c(\Omega)\ TV_\Omega(u), \tag{5.2.7}$$

where $\overline{u}$ is the mean value of u:

$$\overline{u} = \frac{1}{m(\Omega)} \int_\Omega u(x)\, dx.$$

2. Let $\Gamma^0 \subset \partial\Omega$, $m(\Gamma^0) > 0$. There exists a constant $c(\Omega) > 0$ only depending on Ω such that, for all $u \in BV(\Omega)$ satisfying $u = 0$ on Γ^0,

$$\left(\int_\Omega |\, u \,|^{\frac{N}{N-1}}\, dx\right)^{\frac{N-1}{N}} \le c(\Omega)\; TV_\Omega(u). \tag{5.2.8}$$

Actually, the constant $c(\Omega)$ involved in Theorems 5.2.1 and 5.2.2 depends only on θ and r such that the domain Ω has the cone property for these parameters (see [1, Lemma 4-24], [134, Theorem 8-8]).

5.3 Discrete functional inequalities in the general case

We first consider the general case $u \in X(\mathfrak{M})$ with the discrete $W^{1,p}$-norm defined by (5.2.3). The discrete functional inequalities we will prove may be useful in the convergence analysis of finite volume methods for problems with homogeneous Neumann boundary conditions.

5.3.1 General discrete Gagliardo-Nirenberg-Sobolev inequality

We start with the discrete Gagliardo-Nirenberg-Sobolev inequalities which are the discrete counterpart of (5.1.1).

Theorem 5.3.1 (General discrete Gagliardo-Nirenberg-Sobolev inequality)**.** *Let Ω be an open bounded polyhedral domain of $\mathbb{R}^N$, $N \ge 2$. Let $\mathfrak{M}$ be an admissible mesh satisfying (5.2.1)-(5.2.2).*
Then for $p \ge 1$ and $q \ge 1$, there exists a constant $C > 0$ only depending on p, q, θ, N and Ω such that

$$\| u \|_{0,m,\mathfrak{M}} \le \frac{C}{\xi^{(p-1)\theta/p}}\, \| u \|^{\theta}_{1,p,\mathfrak{M}}\, \| u \|^{1-\theta}_{0,q,\mathfrak{M}},\ \forall u \in X(\mathfrak{M}), \tag{5.3.1}$$

where

$$0 \le \theta \le \frac{p}{p + q(p-1)} \le 1 \tag{5.3.2}$$

and

$$\frac{1}{m} = \frac{1-\theta}{q} + \frac{\theta}{p} - \frac{\theta}{N}. \tag{5.3.3}$$

Proof. Throughout this proof, C denotes constants which depend only on Ω, N, p, q and θ. As seen in Section 5.2.2, we have $\|v\|_{BV(\Omega)} \le \|v\|_{1,1,\mathfrak{M}}$ for all $v \in X(\mathfrak{M})$. Therefore, applying Theorem 5.2.1, we get

$$\| v \|_{0,N/(N-1),\mathfrak{M}} \le c(\Omega)\; (|\, v \,|_{1,1,\mathfrak{M}} + \| v \|_{0,1,\mathfrak{M}}) \quad \forall v \in X(\mathfrak{M}). \tag{5.3.4}$$

Let $s \geq 1$. For $u \in X(\mathfrak{M})$, we now define $v \in X(\mathfrak{M})$ by $v_K = |u_K|^s$ for all $K \in \mathfrak{M}$. We note that

$$\| v \|_{0,N/(N-1),\mathfrak{M}} = \left(\sum_{K \in \mathfrak{M}} m(K) |u_K|^{\frac{sN}{N-1}} \right)^{\frac{N-1}{N}} = \| u \|^s_{0,sN/(N-1),\mathfrak{M}},$$
$$\| v \|_{0,1,\mathfrak{M}} = \sum_{K \in \mathfrak{M}} m(K) |u_K|^s = \| u \|^s_{0,s,\mathfrak{M}},$$

and

$$| v |_{1,1,\mathfrak{M}} = \sum_{\substack{\sigma \in \mathcal{E}_{int} \\ \sigma = K|L}} m(\sigma) \Big| |u_K|^s - |u_L|^s \Big|.$$

But, for all $\sigma = K|L$, we have

$$\Big| |u_K|^s - |u_L|^s \Big| \leq s \left(|u_K|^{s-1} + |u_L|^{s-1} \right) |u_K - u_L| .$$

Applying a discrete integration by parts and Hölder's inequality, we get, for any $p \geq 1$ and $s \geq 1$:

$$\sum_{\substack{\sigma \in \mathcal{E}_{int} \\ \sigma = K|L}} m(\sigma) \Big| |u_K|^s - |u_L|^s \Big| \leq s \sum_{K \in \mathfrak{M}} \sum_{\sigma = K|L} \mathrm{m}(\sigma) |u_K|^{s-1} |u_K - u_L|$$
$$\leq s \left(\sum_{K \in \mathfrak{M}} \sum_{\sigma = K|L} \frac{\mathrm{m}(\sigma)}{d_\sigma^{p-1}} |u_L - u_K|^p \right)^{\frac{1}{p}} \left(\sum_{K \in \mathfrak{M}} \sum_{\sigma \in \mathcal{E}_K} \mathrm{m}(\sigma)\, d_\sigma\, |u_K|^{\frac{(s-1)p}{p-1}} \right)^{\frac{p-1}{p}}.$$

But, the regularity constraint (5.2.1) on the mesh ensures that for all $K \in \mathfrak{M}$:

$$\sum_{\sigma \in \mathcal{E}_K} \mathrm{m}(\sigma)\, d_\sigma \leq \frac{1}{\xi} \sum_{\sigma \in \mathcal{E}_K} \mathrm{m}(\sigma)\, \mathrm{d}(x_K, \sigma) = \frac{N}{\xi} \mathrm{m}(K),$$

and then, for any $p > 1$, $s \geq 1$, we get:

$$\sum_{\substack{\sigma \in \mathcal{E}_{int} \\ \sigma = K|L}} m(\sigma) \Big| |u_K|^s - |u_L|^s \Big| \leq \frac{Cs}{\xi^{(p-1)/p}} \, | u |_{1,p,\mathfrak{M}} \, \| u \|^{(s-1)}_{0,(s-1)p/(p-1),\mathfrak{M}}.$$

Therefore, from (5.3.4), we obtain that for all $u \in X(\mathfrak{M})$, $p \geq 1$ and $s \geq 1$:

$$\| u \|^s_{0,sN/(N-1),\mathfrak{M}} \leq C \left(\frac{1}{\xi^{(p-1)/p}} \, | u |_{1,p,\mathfrak{M}} \, \| u \|^{s-1}_{0,(s-1)p/(p-1),\mathfrak{M}} + \|u\|^s_{0,s,\mathfrak{M}} \right).$$

Let $q \geq 1$. If $p = 1$ we choose $s = 1$, and if $p > 1$ we choose $s := 1 + (p-1)q/p > 1$, which yields that $q = (s-1)p/(p-1)$. We obtain by interpolation between L^r spaces that since

$$\frac{1}{s} = \frac{1/s}{p} + \frac{(s-1)/s}{q} \leq 1,$$

$$\| u \|_{0,s,\mathfrak{M}} \leq \| u \|_{0,p,\mathfrak{M}}^{1/s} \| u \|_{0,q,\mathfrak{M}}^{(s-1)/s}, \tag{5.3.5}$$

which yields

$$\| u \|_{0,sN/(N-1),\mathfrak{M}}^{s} \leq C \left(\frac{1}{\xi^{(p-1)/p}} \, | u |_{1,p,\mathfrak{M}} \| u \|_{0,q,\mathfrak{M}}^{s-1} + \| u \|_{0,p,\mathfrak{M}} \| u \|_{0,q,\mathfrak{M}}^{s-1} \right),$$

and finally

$$\| u \|_{0,sN/(N-1),\mathfrak{M}} \leq \frac{C}{\xi^{(p-1)/(ps)}} \| u \|_{0,q,\mathfrak{M}}^{(s-1)/s} \| u \|_{1,p,\mathfrak{M}}^{1/s}, \quad \forall p \geq 1, \quad \forall q \geq 1, \tag{5.3.6}$$

with $s = 1 + (p-1)q/p$.
Let $0 \leq \alpha \leq 1$. We take $m \geq 1$ such that

$$\frac{1}{m} = \frac{\alpha}{sN/(N-1)} + \frac{1-\alpha}{q}.$$

By interpolation, we have:

$$\begin{aligned} \| u \|_{0,m,\mathfrak{M}} &\leq \| u \|_{0,sN/(N-1),\mathfrak{M}}^{\alpha} \| u \|_{0,q,\mathfrak{M}}^{1-\alpha} \\ &\leq \frac{C}{\xi^{\alpha(p-1)/(ps)}} \| u \|_{1,p,\mathfrak{M}}^{\alpha/s} \| u \|_{0,q,\mathfrak{M}}^{1-\alpha/s}, \quad \forall 0 \leq \alpha \leq 1, \quad \forall p \geq 1, \quad \forall q \geq 1, \end{aligned}$$

with

$$s = 1 + \frac{(p-1)q}{p} \quad \text{and} \quad \frac{1}{m} = \frac{\alpha}{sN/(N-1)} + \frac{1-\alpha}{q}.$$

Setting $\theta = \alpha/s$, we get the expected inequality (5.3.1) under the conditions (5.3.2) and (5.3.3). □

5.3.2 General discrete Sobolev-Poincaré inequality

We now give the discrete counterpart of the Sobolev-Poincaré inequalities (5.1.2).

Theorem 5.3.2 (General discrete Sobolev-Poincaré inequality)**.** *Let Ω be an open bounded polyhedral domain of $\mathbb{R}^N$, $N \geq 2$. Let $\mathfrak{M}$ be an admissible mesh satisfying (5.2.1)-(5.2.2). Then there exists a constant $C > 0$ only depending on p, q, N and Ω such that:*

- *if $1 \leq p < N$, for all $1 \leq q \leq p^* := \dfrac{pN}{N-p}$,*

$$\| u \|_{0,q,\mathfrak{M}} \leq \frac{C}{\xi^{(p-1)/p}} \| u \|_{1,p,\mathfrak{M}}, \quad \forall u \in X(\mathfrak{M}), \tag{5.3.7}$$

- *if $p \geq N$, for all $1 \leq q < +\infty$,*

$$\| u \|_{0,q,\mathfrak{M}} \leq \frac{C}{\xi^{(p-1)/p}} \| u \|_{1,p,\mathfrak{M}}, \quad \forall u \in X(\mathfrak{M}).$$

Proof. Throughout this proof, C denotes constants which depend only on Ω, N, p and q. The proof is divided into four steps corresponding to different values of p: the case $p = 1$, the case $1 < p < N$, the critical case $p = N$ and finally the case $p > N$.

Case $p = 1$. If $q = p^* = N/(N-1)$, it corresponds to estimate (5.3.4). Then we obtain (5.3.7) for $p = 1$ and for all $1 \le q \le p^*$ by using the fact that $L^{p^*}(\Omega) \subset L^q(\Omega)$ since Ω is bounded.

Case $1 < p < N$. We get from (5.3.6) that

$$\| u \|_{0,sN/(N-1),\mathfrak{M}} \le \frac{C}{\xi^{(p-1)/(ps)}} \| u \|_{0,q,\mathfrak{M}}^{(s-1)/s} \| u \|_{1,p,\mathfrak{M}}^{1/s}, \quad \forall p > 1, \quad \forall q \ge 1, \tag{5.3.8}$$

with

$$s = 1 + \frac{(p-1)q}{p} \ge 1.$$

Then we choose $q \ge 1$ such that

$$q = \frac{sN}{N-1}, \quad \text{that is} \quad q = \frac{pN}{N-p} \ge 1 \quad \text{for } p < N.$$

Therefore, we get

$$\| u \|_{0,pN/(N-p),\mathfrak{M}} \le \frac{C}{\xi^{(p-1)/p}} \| u \|_{1,p,\mathfrak{M}} \quad \forall 1 < p < N,$$

and since $L^{pN/(N-p)}(\Omega) \subset L^q(\Omega)$ for all $1 \le q \le pN/(N-p)$, it yields

$$\| u \|_{0,q,\mathfrak{M}} \le \frac{C}{\xi^{(p-1)/p}} \| u \|_{1,p,\mathfrak{M}} \quad \forall q \in \left[1 , \frac{pN}{N-p} \right],$$

and the proof is complete for $1 \le p < N$.

Case $p = N$. As in the proof of Theorem 5.3.1, for all $q \ge 1$ we choose $s = 1 + (N-1)q/N \ge 1$, which yields that $q = (s-1)N/(N-1) \ge 1$. Using the Young's inequality in (5.3.6), we obtain

$$\| u \|_{0,sN/(N-1),\mathfrak{M}} \le C_\star \left(\frac{1}{\xi^{(N-1)/N}} \| u \|_{1,N,\mathfrak{M}} + \| u \|_{0,(s-1)N/(N-1),\mathfrak{M}} \right), \tag{5.3.9}$$

with $C_\star$ independant of s. Then we proceed by induction on $s \in \mathbb{N}$, $s \ge N-1$ to prove that there exists a constant C_s depending on s such that

$$\| u \|_{0,sN/(N-1),\mathfrak{M}} \le \frac{C_s}{\xi^{(N-1)/N}} \| u \|_{1,N,\mathfrak{M}}. \tag{5.3.10}$$

For $s = N-1$, the result is already given by (5.3.8). Then let $s \in \mathbb{N}$, $s \ge N$ such that (5.3.10) is true for $s-1$. Using (5.3.9) and (5.3.10), we get

$$\| u \|_{0,sN/(N-1),\mathfrak{M}} \le \frac{C_\star(1 + C_{s-1})}{\xi^{(N-1)/N}} \| u \|_{1,N,\mathfrak{M}}.$$

Then (5.3.10) is true for all integer $s \ge N-1$, and since $L^s(\Omega) \subset L^q(\Omega)$ for all real number $q \le s$, it finally yields that

$$\| u \|_{0,q,\mathfrak{M}} \le \frac{C_q}{\xi^{(N-1)/N}} \| u \|_{1,N,\mathfrak{M}} \quad \forall q \in [1, +\infty[, \tag{5.3.11}$$

which is the result for $p = N$. We emphasize that $C_q \to +\infty$ as $q \to +\infty$.

Case $p > N$. We obtain the result using the fact that

$$\| u \|_{1,N,\mathfrak{M}} \leq \frac{C}{\xi^{(p-N)/pN}} \| u \|_{1,p,\mathfrak{M}} \quad \forall p \geq N. \tag{5.3.12}$$

Gathering (5.3.11) and (5.3.12) we get

$$\| u \|_{0,q,\mathfrak{M}} \leq \frac{C}{\xi^{(N-1)/N}} \| u \|_{1,N,\mathfrak{M}} \leq \frac{C}{\xi^{(p-1)/p}} \| u \|_{1,p,\mathfrak{M}} \quad \forall q \in [1, +\infty[,$$

which completes the proof of Theorem 5.3.2.

□

5.3.3 Other discrete functional inequalities

From Theorems 5.3.1 and 5.3.2, we can deduce a discrete Nash inequality:

Corollary 5.3.1 (Discrete Nash inequality)**.** *Let Ω be an open bounded polyhedral domain of $\mathbb{R}^N$. Let $\mathfrak{M}$ be an admissible mesh satisfying (5.2.1)-(5.2.2). Then there exists a constant $C > 0$ only depending on Ω and N such that*

$$\| u \|_{0,2,\mathfrak{M}}^{1+\frac{2}{N}} \leq \frac{C}{\sqrt{\xi}} \| u \|_{1,2,\mathfrak{M}} \| u \|_{0,1,\mathfrak{M}}^{\frac{2}{N}}, \quad \forall u \in X(\mathfrak{M}).$$

Proof. For $N = 2$, the result is directly given by the application of Theorem 5.3.1 with $p = 2$, $q = 1$, $\theta = 1/N = 1/2$ and $m = 2$. For $N \geq 3$, let us first apply Hölder's inequality:

$$\| u \|_{0,2,\mathfrak{M}}^2 = \sum_{K \in \mathfrak{M}} \mathrm{m}(K) |u_K|^{4/(N+2)} |u_K|^{2N/(N+2)} \leq \| u \|_{0,1,\mathfrak{M}}^{4/(N+2)} \| u \|_{0,2N/(N-2),\mathfrak{M}}^{2N/(N+2)}. \tag{5.3.13}$$

Then we apply Theorem 5.3.2 with $1 \leq p = 2 < N$ and $q = p^* = 2N/(N-2)$:

$$\| u \|_{0,2N/(N-2),\mathfrak{M}} \leq \frac{C}{\sqrt{\xi}} \| u \|_{1,2,\mathfrak{M}}. \tag{5.3.14}$$

Gathering (5.3.13) and (5.3.14), it yields the result.

□

In the proofs of Theorem 5.3.1 and Theorem 5.3.2, we have used the continuous embedding of $BV(\Omega)$ into $L^{N/(N-1)}(\Omega)$ as it is written in Theorem 5.2.1. But, starting with (5.2.7) instead of (5.2.6) leads to the following discrete Poincaré-Wirtinger inequality.

Theorem 5.3.3 (Discrete Poincaré-Wirtinger inequality)**.** *Let Ω be an open bounded connected polyhedral domain of $\mathbb{R}^N$. Let $\mathfrak{M}$ be an admissible mesh satisfying (5.2.1)-(5.2.2). Then for $1 \leq p < +\infty$ there exists a constant $C > 0$ only depending on Ω, N and p such that:*

$$\| u - \overline{u} \|_{0,p,\mathfrak{M}} \leq \frac{C}{\xi^{(p-1)/p}} | u |_{1,p,\mathfrak{M}} \quad \forall u \in X(\mathfrak{M}). \tag{5.3.15}$$

We recall that $\bar{u} = \dfrac{1}{m(\Omega)} \displaystyle\int_\Omega u(x)dx = \frac{1}{m(\Omega)} \sum_{K \in \mathfrak{M}} m(K) u_K$ for $u \in X(\mathfrak{M})$.

Proof. Throughout this proof, C denotes constants which depend only on Ω, N and p. Using the Hölder's inequality and (5.2.7), we have for all $u \in X(\mathfrak{M})$:

$$\| u - \overline{u} \|_{0,p,\mathfrak{M}} \leq C \| u - \overline{u} \|_{0,N/(N-1),\mathfrak{M}} \leq C \, | u |_{1,1,\mathfrak{M}} \quad \forall 1 \leq p \leq \frac{N}{N-1}.$$

Then applying again the Hölder's inequality, we also have

$$| u |_{1,1,\mathfrak{M}} \leq \frac{C}{\xi^{(p-1)/p}} \, | u |_{1,p,\mathfrak{M}}.$$

Gathering the two latter inequalities, it gives the result (5.3.15) for $1 \leq p \leq N/(N-1)$:

$$\| u - \overline{u} \|_{0,p,\mathfrak{M}} \leq \frac{C}{\xi^{(p-1)/p}} \, | u |_{1,p,\mathfrak{M}} \quad \forall 1 \leq p \leq \frac{N}{N-1}. \tag{5.3.16}$$

Now let us take $s \geq 1$ and for $u \in X(\mathfrak{M})$, we define $v \in X(\mathfrak{M})$ by $v_K = | u_K - \overline{u} |^s$ for all $K \in \mathfrak{M}$. On the one hand, we have

$$\| v - \overline{v} \|_{0,N/(N-1),\mathfrak{M}} \geq \| v \|_{0,N/(N-1),\mathfrak{M}} - \| \overline{v} \|_{0,N/(N-1),\mathfrak{M}}$$

$$\geq \| u - \overline{u} \|^s_{0,sN/(N-1),\mathfrak{M}} - \frac{\| u - \overline{u} \|^s_{0,s,\mathfrak{M}}}{\mathrm{m}(\Omega)^{1/N}}. \tag{5.3.17}$$

On the other hand, using the same technique as in the proof of Theorem 5.3.1, we have:

$$| v |_{1,1,\mathfrak{M}} \leq \frac{C}{\xi^{(p-1)/p}} \, | u |_{1,p,\mathfrak{M}} \| u - \overline{u} \|^{(s-1)}_{0,(s-1)p/(p-1),\mathfrak{M}} \quad \forall p > 1, \quad \forall s \geq 1. \tag{5.3.18}$$

Therefore, gathering (5.3.17) and (5.3.18), we get from (5.2.7) that for $u \in X(\mathfrak{M})$, $p > 1$ and $s \geq 1$:

$$\| u - \overline{u} \|^s_{0,sN/(N-1),\mathfrak{M}} \leq \frac{Cs}{\xi^{(p-1)/p}} \, | u |_{1,p,\mathfrak{M}} \| u - \overline{u} \|^{(s-1)}_{0,(s-1)p/(p-1),\mathfrak{M}} + \frac{\| u - \overline{u} \|^s_{0,s,\mathfrak{M}}}{\mathrm{m}(\Omega)^{1/N}}.$$

Moreover, by interpolation between L^r spaces, for $q \geq 1$, $p > 1$, $s \geq 1$ such that

$$\frac{1}{s} = \frac{1/s}{p} + \frac{(s-1)/s}{q}$$

we have that

$$\| u - \overline{u} \|_{0,s,\mathfrak{M}} \leq \| u - \overline{u} \|^{1/s}_{0,p,\mathfrak{M}} \| u - \overline{u} \|^{(s-1)/s}_{0,q,\mathfrak{M}}.$$

Choosing $q = sN/(N-1) \geq 1$ and $s = (N-1)p/(N-p)$ for $1 < p < N$, it yields

$$\| u - \overline{u} \|_{0,q,\mathfrak{M}} \leq C \left(\frac{1}{\xi^{(p-1)/p}} \, | u |_{1,p,\mathfrak{M}} + \| u - \overline{u} \|_{0,p,\mathfrak{M}} \right) \quad \forall 1 \leq p < N, \tag{5.3.19}$$

where

$$q = q(p) := \frac{pN}{N-p}.$$

Applying (5.3.16) and using the fact that since $p < q$, the Hölder's inequality provides:

$$| u |_{1,p,\mathfrak{M}} \leq \frac{C}{\xi^{1/N}} | u |_{1,q,\mathfrak{M}}, \tag{5.3.20}$$

we can estimate the right hand side of (5.3.19) for $1 \leq p \leq N/(N-1)$, which yields:

$$\| u - \overline{u} \|_{0,q,\mathfrak{M}} \leq \frac{C}{\xi^{(q-1)/q}} | u |_{1,q,\mathfrak{M}} \quad \forall 1 \leq q \leq \frac{N}{N-2}.$$

Using exactly the same technique, we proceed by induction to prove the result for $1 \leq p \leq N/(N-k)$, up to $k = N-1$, which finally yields the result for all $1 \leq p \leq N$.
To conclude, we apply (5.3.19). Indeed, using the result for $1 \leq p < N$ and the inequality (5.3.20), since $q = q(p) = pN/(N-p) \in [1; +\infty)$ if $p \in [1; N)$, we obtain the general result:

$$\| u - \overline{u} \|_{0,q,\mathfrak{M}} \leq \frac{C}{\xi^{(q-1)/q}} | u |_{1,q,\mathfrak{M}} \quad \forall 1 \leq q < +\infty.$$

□

5.4 Discrete functional inequalities in the case of Dirichlet boundary conditions

In this section, we consider the case where the finite volume approximation $u \in X(\mathfrak{M})$ is coming from a finite volume scheme where homogeneous boundary conditions are prescribed on a part of the boundary. This part of the boundary is denoted by $\Gamma^0 \subset \Gamma$, $\mathrm{m}(\Gamma^0) > 0$. In this case, the natural discrete counterparts of the $W^{1,p}$-seminorm and $W^{1,p}$-norm are defined by (5.2.4) and (5.2.5). Moreover, the $W^{1,p}$-seminorm becomes a norm on the space of $W^{1,p}$ functions vanishing on a part of the boundary and the Gagliardo-Nirenberg-Sobolev inequalities and the Sobolev-Poincaré inequalities may be rewritten with the $W^{1,p}$-seminorm instead of the $W^{1,p}$-norm. Our aim in this section is to prove the discrete counterpart of such inequalities (see Theorem 5.4.1 and Theorem 5.4.2).

As in the general case, the starting point will be the continuous embedding from $BV(\Omega)$ into $L^{N/(N-1)}(\Omega)$, which rewrites as (5.2.8) with homogeneous Dirichlet boundary conditions on the part of the boundary. However, (5.2.8) can not be directly applied to $u \in X(\mathfrak{M})$. Indeed, $u \in X(\mathfrak{M})$ belongs to $BV(\Omega)$ and therefore its trace on the boundary is well defined; but it does not necessarily vanish on Γ^0. Some adaptations must be done in order to apply (5.2.8) and get its discrete counterpart. It will be done in Section 5.4.1 and yield the discrete functional inequalities presented in Section 5.4.2 and Section 5.4.3.

In this section, we assume the open set Ω is also convex. This will be particularly crucial in Lemma 5.4.1.

5.4.1 Preliminary lemma

We begin with a lemma which gives the discrete counterpart of (5.2.8). This lemma is crucial to prove Theorems 5.4.1 and 5.4.2.

Lemma 5.4.1. *Let Ω be an open convex bounded polyhedral domain of $\mathbb{R}^N$ and Γ^0 be a part of the boundary Γ such that $m(\Gamma^0) > 0$. Let $\mathfrak{M}$ be an admissible mesh satisfying (5.2.1)-(5.2.2). Then there exists a constant $c(\Omega)$ only depending on Ω such that*

$$\| u \|_{0,N/(N-1),\mathfrak{M}} \leq c(\Omega) \, | u |_{1,1,\Gamma^0,\mathfrak{M}}, \quad \forall u \in X(\mathfrak{M}).$$

Proof. Let us consider $u \in X(\mathfrak{M})$; since u is piecewise constant, u belongs to $BV(\Omega)$. Then we can define the trace Tu of u by: for almost every $x \in \Gamma$,

$$\lim_{r \to 0} \frac{1}{\mathrm{m}\,(B(x,r) \cap \Omega)} \int_{B(x,r) \cap \Omega} |u - Tu(x)| \, dy = 0.$$

Thus in general $Tu_{|\Gamma^0} \neq 0$ and in this framework we cannot take into account the prescribed homogeneous Dirichlet boundary conditions $u_\sigma = 0$ for $\sigma \in \mathcal{E}_{ext}^0$. Therefore the idea is to thicken the domain Ω into a larger domain Ω_ε with $\Omega \subset \Omega_\varepsilon$ and define an extension u_ε of u to Ω_ε such that $u_\varepsilon \in BV(\Omega_\varepsilon)$ and $Tu_\varepsilon = 0$ on a part of positive measure of the boundary $\partial\Omega_\varepsilon$, which allows to apply Theorem 5.2.2 to u_ε.
Let $\sigma \in \mathcal{E}_{ext}$ be a face included in the boundary Γ. Then σ is a part of an hyperplane $\mathcal{H}$ in $\mathbb{R}^N$. We denote by $\mathbf{n}_\sigma$ the unit vector normal to $\mathcal{H}$ outward to Ω. For every $x \in \mathbb{R}^N$, there exists a unique $(y, y') \in \mathcal{H} \times \mathbb{R}$ such that $x = y + y' \, \mathbf{n}_\sigma$. For $\varepsilon > 0$, we define (see Figure 5.1)

$$K_\sigma^\varepsilon := \left\{ x = y + y' \, \mathbf{n}_\sigma \in \mathbb{R}^N \, : \, y \in \sigma \text{ and } 0 < y' < \varepsilon \right\}.$$

Since Ω is convex, $K_\sigma^\varepsilon \cap K_{\sigma'}^\varepsilon$ is empty for all σ, $\sigma' \in \mathcal{E}_{ext}$ with $\sigma \neq \sigma'$. Now we can define (see Figure 5.1)

$$\Omega_\varepsilon := \Omega \cup \left(\bigcup_{\sigma \in \mathcal{E}_{ext}} K_\sigma^\varepsilon \right).$$

The subset Ω_ε is polyhedral, then it is a Lipschitz domain. We point out that for all $x \in \Omega_\varepsilon \setminus \Omega$, $\mathrm{d}(x, \Omega) \leq \varepsilon$ and then if we consider a face between two new cells K_σ^ε and $K_{\sigma'}^\varepsilon$, we have:

$$\mathrm{m}\left(\overline{K_\sigma^\varepsilon} \cap \overline{K_{\sigma'}^\varepsilon} \right) \leq C \, \varepsilon^{N-1}.$$

Then we define

$$u_{K_\sigma^\varepsilon} := \begin{cases} 0 & \text{if} \quad \sigma \in \mathcal{E}_{ext}^0, \\ u_K & \text{if} \quad \sigma \in \mathcal{E}_{ext} \setminus \mathcal{E}_{ext}^0, \quad \sigma \in \mathcal{E}_K, \end{cases}$$

and the function u_ε in the following way:

$$u_\varepsilon := \sum_{K \in \mathfrak{M}} u_K \mathbf{1}_K + \sum_{\sigma \in \mathcal{E}_{ext}} u_{K_\sigma^\varepsilon} \mathbf{1}_{K_\sigma^\varepsilon}.$$

We have obviously

$$\| u \|_{0,N/(N-1),\mathfrak{M}} \leq \| u_\varepsilon \|_{L^{N/(N-1)}(\Omega_\varepsilon)}.$$

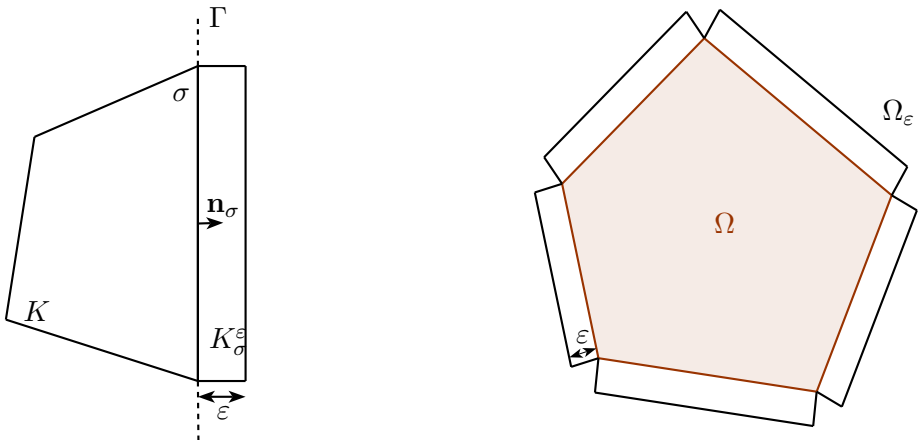


Figure 5.1: Construction of the cell K_σ^ε (left) and of the domain Ω_ε (right).

Moreover, since the function u_ε is piecewise constant and has a finite number of jumps (which corresponds to the number of faces $\sigma \in \mathcal{E}$ added to the number of faces between the new cells K_σ^ε) we get that u_ε belongs to $BV(\Omega_\varepsilon)$, and

$$TV_{\Omega_\varepsilon}(u_\varepsilon) \leq \sum_{\sigma\in\mathcal{E}} \mathrm{m}(\sigma)|D_\sigma u| + \sum_{\overline{K_\sigma^\varepsilon}\cap\overline{K_{\sigma'}^\varepsilon}\neq\emptyset} C\,\varepsilon^{N-1}\left|u_{K_\sigma^\varepsilon} - u_{K_{\sigma'}^\varepsilon}\right|.$$

Furthermore, since $u_\varepsilon = 0$ on a part of positive measure of the boundary $\partial\Omega_\varepsilon$, we can apply the result (5.2.8) of Theorem 5.2.2 to u_ε. We obtain that there exists a constant $c(\Omega_\varepsilon)$ such that

$$\|\,u\,\|_{0,N/(N-1),\mathfrak{M}} \;\leq\; c(\Omega_\varepsilon)\left(|\,u\,|_{1,1,\Gamma^0,\mathfrak{M}} + C\,\varepsilon^{N-1}\sum_{\overline{K_\sigma^\varepsilon}\cap\overline{K_{\sigma'}^\varepsilon}\neq\emptyset}\left|u_{K_\sigma^\varepsilon} - u_{K_{\sigma'}^\varepsilon}\right|\right).$$

Now since the open set Ω is polyhedral, it is a Lipschitz domain then it satisfies the cone condition [1, Definition 4-6] for some cone C and by construction the open set Ω_ε also satisfies the cone condition with the same cone. Therefore, applying Lemma 4-24 in [1], the constant $c(\Omega_\varepsilon)$ only depends on the dimension of this cone, and not on $\varepsilon > 0$. Then passing to the limit $\varepsilon \to 0$ we finally get that

$$\|\,u\,\|_{0,N/(N-1),\mathfrak{M}} \;\leq\; c(\Omega)\,|\,u\,|_{1,1,\mathfrak{M}}.$$

□

Now using this lemma we can prove the discrete Gagliardo-Nirenberg-Sobolev and Sobolev-Poincaré inequalities in the case with some homogeneous Dirichlet boundary conditions.

5.4.2 Discrete Gagliardo-Nirenberg-Sobolev inequality

Theorem 5.4.1 (Discrete Gagliardo-Nirenberg-Sobolev inequality)**.** *Let Ω be an open convex bounded polyhedral domain of $\mathbb{R}^N$ and Γ^0 be a part of the boundary such that $m(\Gamma^0) > 0$. Let $\mathfrak{M}$ be an admissible mesh satisfying (5.2.1)-(5.2.2). Then for any $p \geq 1$ and $q \geq 1$, there exists a constant $C > 0$ only depending on p, q, N, θ and Ω such that*

$$\| u \|_{0,m,\mathfrak{M}} \leq \frac{C_1}{\xi^{(p-1)\theta/p}} \, | u |^{\theta}_{1,p,\Gamma^0,\mathfrak{M}} \, \| u \|^{1-\theta}_{0,q,\mathfrak{M}}, \quad \forall u \in X(\mathfrak{M}),$$

where θ and m satisfy (5.3.2) *and* (5.3.3)*.*

Proof. The proof is similar to the proof of Theorem 5.3.1. Let $p \geq 1$ and $s \geq 1$. For $u \in X(\mathfrak{M})$, we apply Lemma 5.4.1 to $v \in X(\mathfrak{M})$ defined by $v_K = u_K^s$ for all $K \in \mathfrak{M}$. It yields

$$\|u\|^s_{0,sN/(N-1),\mathfrak{M}} \leq \frac{C}{\xi^{(p-1)/p}} \, | u |_{1,p,\Gamma^0,\mathfrak{M}} \, \| u \|^{(s-1)}_{0,(s-1)p/(p-1),\mathfrak{M}}$$

with $p \geq 1$ and $s \geq 1$.
For $q \geq 1$, choosing $s = 1 + (p-1)q/p \geq 1$, we obtain

$$\| u \|_{0,sN/(N-1),\mathfrak{M}} \leq \frac{C}{\xi^{(p-1)/ps}} \, | u |^{1/s}_{1,p,\Gamma^0,\mathfrak{M}} \, \| u \|^{(s-1)/s}_{0,q,\mathfrak{M}}. \tag{5.4.1}$$

Then, using an interpolation inequality as in the proof of Theorem 5.3.1, we get

$$\| u \|_{0,m,\mathfrak{M}} \leq \| u \|^{\alpha}_{0,sN/(N-1),\mathfrak{M}} \, \| u \|^{1-\alpha}_{0,q,\mathfrak{M}} \leq \frac{C}{\xi^{\alpha(p-1)/ps}} \, | u |^{\alpha/s}_{1,p,\Gamma^0,\mathfrak{M}} \, \| u \|^{1-(\alpha/s)}_{0,q,\mathfrak{M}},$$

Taking $\theta = \alpha/s$ concludes the proof.

□

5.4.3 Discrete Sobolev-Poincaré and Nash inequalities

In the case with some homogeneous Dirichlet boundary conditions, the discrete Sobolev-Poincaré inequalities rewrite as follows.

Theorem 5.4.2 (Discrete Sobolev-Poincaré inequality)**.** *Let Ω be an open convex bounded polyhedral domain of $\mathbb{R}^N$. Let $\mathfrak{M}$ be an admissible mesh satisfying (5.2.1)-(5.2.2). Let $\Gamma^0 \subset \Gamma$, with $m(\Gamma^0) > 0$.*
Then there exists a constant $C > 0$ which only depends on p, q, N and Ω such that:

- *if $1 \leq p < N$, for all $1 \leq q \leq p^* := \dfrac{pN}{N-p}$*

$$\| u \|_{0,q,\mathfrak{M}} \leq \frac{C}{\xi^{(p-1)/p}} \, | u |_{1,p,\Gamma^0,\mathfrak{M}} \quad \forall u \in X(\mathfrak{M})$$

- *if $p \geq N$, for all $1 \leq q < +\infty$,*

$$\| u \|_{0,q,\mathfrak{M}} \leq \frac{C}{\xi^{(p-1)/p}} \, | u |_{1,p,\Gamma^0,\mathfrak{M}} \quad \forall u \in X(\mathfrak{M}).$$

Proof. The proof is similar to the proof of Theorem 5.3.2, starting from (5.4.1) instead of (5.3.6). □

Now using Theorems 5.4.1 and 5.4.2, we easily get a discrete version of Nash inequality:

Corollary 5.4.1 (Discrete Nash inequality)**.** *Let Ω be an open convex bounded polyhedral domain of $\mathbb{R}^N$. Let $\mathfrak{M}$ be an admissible mesh satisfying (5.2.1)-(5.2.2). Let $\Gamma^0 \subset \Gamma$, with $m(\Gamma^0) > 0$.*
Then there exists a constant $C > 0$ only depending on Ω and N such that

$$\| u \|_{0,2,\mathfrak{M}}^{1+\frac{2}{N}} \leq \frac{C}{\sqrt{\xi}} \, | u |_{1,2,\Gamma^0,\mathfrak{M}} \, \| u \|_{0,1,\mathfrak{M}}^{\frac{2}{N}} \quad \forall u \in X(\mathfrak{M}).$$

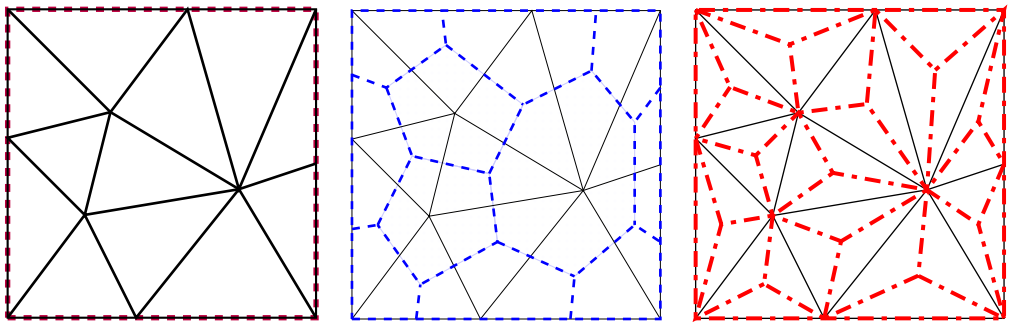
5.5 Application to finite volume approximations coming from DDFV schemes

The discrete duality finite volume methods have been developed for ten years for the approximation of anisotropic elliptic problems on almost general meshes in 2D and 3D. They are based on some discrete operators (divergence and gradient), satisfying a discrete Green formula (the "discrete duality"). The DDFV approximations were first proposed for the discretization of anisotropic and/or nonlinear diffusion problems on rather general meshes. We refer to the pioneer work of F. Hermeline [103, 104, 105, 106, 107] who proposed a new approach dealing with primal and dual meshes and Y. Coudière, J.-P. Vila anf P. Villedieu [63] who proposed a method of reconstruction for the discrete gradients. Next, K. Domelevo and P. Omnès [73], S. Delcourte, K. Domelevo and P. Omnès [69] presented the discrete duality finite volume approach (DDFV) for the Laplace operator. Then, B. Andreianov, F. Boyer and F. Hubert [5] gave a general background of DDFV methods for anisotropic and nonlinear elliptic problems. Most of these works treat 2D linear anisotropic, heterogeneous diffusion problems, while the case of discontinuous diffusion operators have been treated later by F. Boyer and F. Hubert in [27]. F. Hermeline [106, 107] treats the analogous 3D problems, S. Krell [129] treats the Stokes problem in 2D and in 3D whereas Y. Coudière and G. Manzini [61] treat linear elliptic convection-diffusion equations.

The construction of DDFV schemes needs the definition of three meshes: a primal mesh, a dual mesh and a diamond mesh. Then, the approximate solutions are defined both on the primal and the dual meshes, while the approximate gradients are defined on the diamond mesh. Therefore, we need to adapt the definition of the spaces of approximate solutions and the definition of the discrete norms. It will be done in Section 5.5.1. Then, we will be able to establish some discrete Gagliardo-Nirenberg-Sobolev and Sobolev-Poincaré inequalities, in the general case (Section 5.5.2) as in the case with Dirichlet boundary conditions (Section 5.5.3) .

5.5.1 Meshes and functional spaces

Meshes. Let Ω be an open bounded polygonal domain of $\mathbb{R}^2$. The mesh construction starts with the partition of Ω with disjoint open polygonal control volumes. This partition, denoted by $\mathfrak{M}$, is called the interior primal mesh. We then denote by $\partial\mathfrak{M}$ the set of boundary edges, which are considered as degenerate control volumes. Then, the primal mesh is defined by $\overline{\mathfrak{M}} = \mathfrak{M} \cup \partial\mathfrak{M}$. To each primal cell $\mathcal{K} \in \overline{\mathfrak{M}}$, we associate a point $x_\mathcal{K} \in \mathcal{K}$, called the center of the primal cell. Notice that for a degenerate control volume $\mathcal{K}$, the point $x_\mathcal{K}$ is necessarily the midpoint of $\mathcal{K}$. This family of centers is denoted by $\mathcal{X} = \{x_\mathcal{K}, \mathcal{K} \in \overline{\mathfrak{M}}\}$.



Interior and exterior primal meshes | Interior and exterior dual meshes | Diamond mesh

Figure 5.2: Presentation of the meshes

Let $\mathcal{X}^*$ denote the set of the vertices of the primal control volumes in $\overline{\mathfrak{M}}$. Distinguishing the interior vertices from the vertices lying on the boundary, we split $\mathcal{X}^*$ into $\mathcal{X}^* = \mathcal{X}^*_{int} \cup \mathcal{X}^*_{ext}$. To any point $x_{\mathcal{K}^*} \in \mathcal{X}^*_{int}$, we associate the polygon $\mathcal{K}^*$ obtained by joining the centers of the primal cells whose $x_{\mathcal{K}^*}$ is a vertex. The set of such polygons defines the interior dual mesh denoted by $\mathfrak{M}^*$. To any point $x_{\mathcal{K}^*} \in \mathcal{X}^*_{ext}$, we then associate the polygon $\mathcal{K}^*$, whose vertices are $\{x_{\mathcal{K}^*}\} \cup \{x_\mathcal{K} \in \mathcal{X} / x_{\mathcal{K}^*} \in \bar{\mathcal{K}}, \mathcal{K} \in \overline{\mathfrak{M}}\}$. It defines the boundary dual mesh $\partial\mathfrak{M}^*$ and the dual mesh is defined by $\overline{\mathfrak{M}^*} = \mathfrak{M}^* \cup \partial\mathfrak{M}^*$.

In the sequel, we will assume that each primal cell $\mathcal{K} \in \mathfrak{M}$ is star-shaped with respect to $x_\mathcal{K}$ and each dual cell $\mathcal{K}^* \in \overline{\mathfrak{M}^*}$ is star-shaped with respect to $x_{\mathcal{K}^*}$.

For all neighboring primal cells $\mathcal{K}$ and $\mathcal{L}$, we assume that $\partial\mathcal{K} \cap \partial\mathcal{L}$ is a segment, corresponding to an edge of the mesh $\mathfrak{M}$, denoted by $\sigma = \mathcal{K}|\mathcal{L}$. Let $\mathcal{E}$ be the set of such edges. We similarly define the edges $\mathcal{E}^*$ of the dual mesh $\overline{\mathfrak{M}^*}$: $\sigma^* = \mathcal{K}^*|\mathcal{L}^*$. For each couple $(\sigma, \sigma^*) \in \mathcal{E} \times \mathcal{E}^*$ such that $\sigma = \mathcal{K}|\mathcal{L} = (x_{\mathcal{K}^*}, x_{\mathcal{L}^*})$ and $\sigma^* = \mathcal{K}^*|\mathcal{L}^* = (x_\mathcal{K}, x_\mathcal{L})$, we define the quadrilateral diamond cell $\mathcal{D}_{\sigma,\sigma^*}$ whose diagonals are σ and σ^*. If $\sigma \in \mathcal{E} \cap \partial\Omega$, we note that the diamond degenerates into a triangle. The set of the diamond cells defines a partition of Ω, which is called the diamond mesh and is denoted by $\mathfrak{D}$. Let us note that $\mathfrak{D}$ can be splitted into $\mathfrak{D} = \mathfrak{D}_{int} \cup \mathfrak{D}_{ext}$ where $\mathfrak{D}_{int}$ is the set of interior (non degenerate) diamond cells and $\mathfrak{D}_{ext}$ is the set of degenerate diamond cells.

Finally, the DDFV mesh is made of the triple $\mathcal{T} = (\overline{\mathfrak{M}}, \overline{\mathfrak{M}^*}, \mathfrak{D})$. See Figure 5.2 for an

example of DDFV mesh.

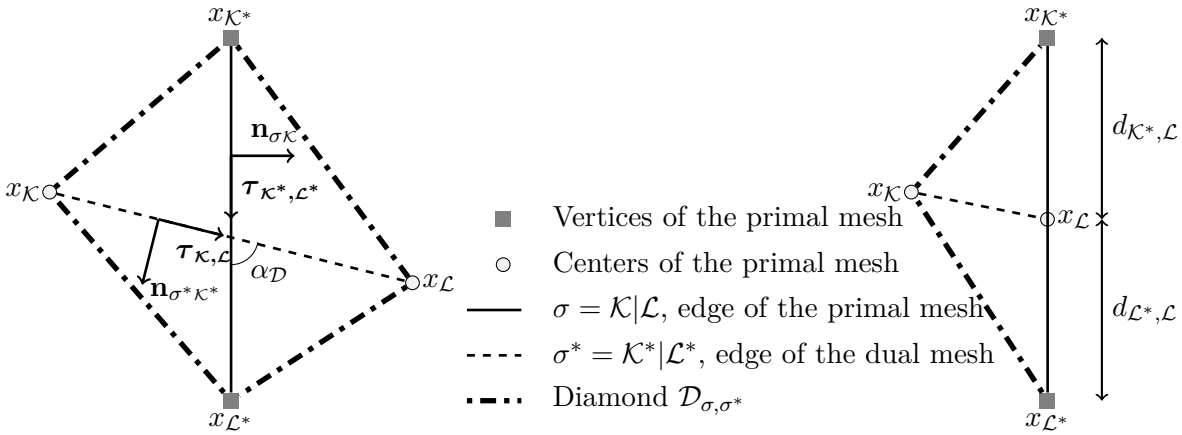


Figure 5.3: Definition of the diamonds $\mathcal{D}_{\sigma,\sigma^*}$

Let us now introduce some notations associated to the mesh $\mathcal{T}$. For each primal cell or dual cell V in $\overline{\mathfrak{M}}$ or $\overline{\mathfrak{M}^*}$, we define m_V, the measure of the cell V, $\mathcal{E}_V$, the set of edges of V, $\mathfrak{D}_V = \{\mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D},\ \sigma \in \mathcal{E}_V\}$, d_V, the diameter of V. For a diamond $\mathcal{D}_{\sigma,\sigma^*}$, whose vertices are $(x_{\mathcal{K}}, x_{\mathcal{K}^*}, x_{\mathcal{L}}, x_{\mathcal{L}^*})$, we define : m_σ and m_{σ^*} the lengths of the primal edge σ and the dual edge σ^*, $\mathrm{m}_{\mathcal{D}}$, the measure of $\mathcal{D}$, $d_{\mathcal{D}}$ its diameter and $\alpha_{\mathcal{D}}$ the angle between $(x_{\mathcal{K}}, x_{\mathcal{L}})$ and $(x_{\mathcal{K}^*}, x_{\mathcal{L}^*})$. As shown on Figure 5.3, we will also use two direct basis $(\boldsymbol{\tau}_{\boldsymbol{\mathcal{K}^*,\mathcal{L}^*}}, \mathbf{n}_{\sigma\mathcal{K}})$ and $(\mathbf{n}_{\sigma^*\mathcal{K}^*}, \boldsymbol{\tau}_{\boldsymbol{\mathcal{K},\mathcal{L}}})$, where $\mathbf{n}_{\sigma\mathcal{K}}$ is the unit normal to σ, outward $\mathcal{K}$, $\mathbf{n}_{\sigma^*\mathcal{K}^*}$ is the unit normal to σ^*, outward $\mathcal{K}^*$, $\boldsymbol{\tau}_{\boldsymbol{\mathcal{K}^*,\mathcal{L}^*}}$ is the unit tangent vector to σ, oriented from $\mathcal{K}^*$ to $\mathcal{L}^*$, $\boldsymbol{\tau}_{\boldsymbol{\mathcal{K},\mathcal{L}}}$ is the unit tangent vector to σ^*, oriented from $\mathcal{K}$ to $\mathcal{L}$. For a boundary edge $\sigma = [x_{\mathcal{K}^*}, x_{\mathcal{L}^*}] \in \partial\mathfrak{M}$, we define $d_{\mathcal{K}^*,\mathcal{L}}$ the length of the segment $[x_{\mathcal{K}^*}, x_{\mathcal{L}}]$ and $d_{\mathcal{L}^*,\mathcal{L}}$ the length of the segment $[x_{\mathcal{L}^*}, x_{\mathcal{L}}]$.

In all the sequel, we will assume that the diamonds cannot be flat. It means :

$$\exists \alpha_{\mathcal{T}} \in]0, \frac{\pi}{2}] \text{ such that } |\sin(\alpha_{\mathcal{D}})| \geq \sin(\alpha_{\mathcal{T}}) \quad \forall \mathcal{D} \in \mathfrak{D}. \tag{5.5.1}$$

As for all $\mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}$, we have $2\mathrm{m}_{\mathcal{D}} = \mathrm{m}_\sigma \mathrm{m}_{\sigma^*} \sin(\alpha_{\mathcal{D}})$, the hypothesis (5.5.1) implies

$$\mathrm{m}_\sigma \mathrm{m}_{\sigma^*} \leq \frac{2\mathrm{m}_{\mathcal{D}}}{\sin(\alpha_{\mathcal{T}})}.$$

We also assume some regularity of the mesh, as in [5], which implies

$$\begin{aligned} &\exists \zeta > 0, \sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}_{\mathcal{K}}} \mathrm{m}_\sigma \mathrm{m}_{\sigma^*} \leq \frac{\mathrm{m}_{\mathcal{K}}}{\zeta} \quad \forall \mathcal{K} \in \mathfrak{M}, \\ &\sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}_{\mathcal{K}^*}} \mathrm{m}_\sigma \mathrm{m}_{\sigma^*} \leq \frac{\mathrm{m}_{\mathcal{K}^*}}{\zeta} \quad \forall \mathcal{K}^* \in \overline{\mathfrak{M}^*}. \end{aligned} \tag{5.5.2}$$

Definition of the approximate solution. A discrete duality finite volume scheme leads to the computation of discrete unknowns on the primal and the dual meshes : $(u_\mathcal{K})_{\mathcal{K}\in\overline{\mathfrak{M}}}$ and $(u_{\mathcal{K}^*})_{\mathcal{K}^*\in\overline{\mathfrak{M}^*}}$. From these discrete unknowns, we can reconstruct two different approximate solutions :

$$u_\mathfrak{M} = \sum_{\mathcal{K}\in\mathfrak{M}} u_\mathcal{K}\mathbf{1}_\mathcal{K} \text{ and } u_{\overline{\mathfrak{M}^*}} = \sum_{\mathcal{K}^*\in\overline{\mathfrak{M}^*}} u_{\mathcal{K}^*}\mathbf{1}_{\mathcal{K}^*}.$$

But, in order to use simultaneously the discrete unknowns computed on the primal and the dual meshes, we prefer to define the approximate solution as

$$u = \frac{1}{2}(u_\mathfrak{M} + u_{\overline{\mathfrak{M}^*}}).$$

Therefore, the space of approximate solutions $Z(\mathcal{T})$ is defined by:

$$Z(\mathcal{T}) = \left\{ u\in L^1(\Omega) \ / \ \exists u_\mathcal{T} = \left((u_\mathcal{K})_{\mathcal{K}\in\overline{\mathfrak{M}}}, (u_{\mathcal{K}^*})_{\mathcal{K}^*\in\overline{\mathfrak{M}^*}}\right) \right.$$
$$\left. \text{such that } u = \frac{1}{2}\left(\sum_{\mathcal{K}\in\mathfrak{M}} u_\mathcal{K}\mathbf{1}_\mathcal{K} + \sum_{\mathcal{K}^*\in\overline{\mathfrak{M}^*}} u_{\mathcal{K}^*}\mathbf{1}_{\mathcal{K}^*}\right)\right\}.$$

For a given function $u\in Z(\mathcal{T})$, we define the discrete L^p-norm by

$$\|u\|_{0,p,\mathcal{T}} = \left(\frac{1}{2}\sum_{\mathcal{K}\in\mathfrak{M}} \mathrm{m}_\mathcal{K}|u_\mathcal{K}|^p + \frac{1}{2}\sum_{\mathcal{K}^*\in\overline{\mathfrak{M}^*}} \mathrm{m}_{\mathcal{K}^*}|u_{\mathcal{K}^*}|^p\right)^{1/p}.$$

Discrete gradient. A key point in the construction of the DDFV schemes is the definition of the discrete operators (divergence and gradient). We just focus here on the definition of the discrete gradient, which will be useful for the definition of the discrete $W^{1,p}$-seminorms.

Let $u\in Z(\mathcal{T})$. The discrete gradient of u, $\nabla^d u$ is defined as a piecewise constant function on each diamond cell :

$$\nabla^d u = \sum_{\mathcal{D}\in\mathfrak{D}} \nabla^\mathcal{D} u\, \mathbf{1}_\mathcal{D},$$

where, for $\mathcal{D}\in\mathfrak{D}$,

$$\nabla^\mathcal{D} u = \frac{1}{\sin(\alpha_\mathcal{D})}\left(\frac{u_\mathcal{L} - u_\mathcal{K}}{\mathrm{m}_{\sigma^*}}\mathbf{n}_{\sigma\mathcal{K}} + \frac{u_{\mathcal{L}^*} - u_{\mathcal{K}^*}}{\mathrm{m}_\sigma}\mathbf{n}_{\sigma^*\mathcal{K}^*}\right).$$

This discrete gradient has been introduced in [63]. It verifies:

$$\nabla^\mathcal{D} u\cdot\boldsymbol{\tau}_{\mathcal{K}^*,\mathcal{L}^*} = \frac{u_{\mathcal{L}^*} - u_{\mathcal{K}^*}}{\mathrm{m}_\sigma} \text{ and } \nabla^\mathcal{D} u\cdot\boldsymbol{\tau}_{\mathcal{K},\mathcal{L}} = \frac{u_\mathcal{L} - u_\mathcal{K}}{\mathrm{m}_{\sigma^*}}.$$

Using this discrete gradient, we may now define the discrete $W^{1,p}$-seminorm and norm of a given function $u \in Z(\mathcal{T})$:

$$\begin{aligned} |u|_{1,p,\mathcal{T}} &= \left(\sum_{\mathcal{D} \in \mathfrak{D}} \mathrm{m}_{\mathcal{D}} |\nabla^{\mathcal{D}} u|^p \right)^{1/p}, \\ \|u\|_{1,p,\mathcal{T}} &= \|u\|_{0,p,\mathcal{T}} + |u|_{1,p,\mathcal{T}}. \end{aligned}$$

5.5.2 Discrete functional inequalities in the general case

Our aim is now to extend the results of Section 5.3 to the case of finite volume approximations coming from some DDFV schemes: $u \in Z(\mathcal{T})$. We will use that such functions are defined as $u = \frac{1}{2}(u_{\mathfrak{M}} + u_{\overline{\mathfrak{M}^*}})$ with $u_{\mathfrak{M}} \in X(\mathfrak{M})$ and $u_{\overline{\mathfrak{M}^*}} \in X(\overline{\mathfrak{M}^*})$. Nevertheless, we may take care because the primal and the dual meshes $\mathfrak{M}$ and $\overline{\mathfrak{M}^*}$ does not satisfy the admissibility condition required in Theorems 5.3.1 and 5.3.2.

Theorem 5.5.1 (General discrete Gagliardo-Nirenberg-Sobolev inequality in the DDFV framework)**.** *Let Ω be an open bounded polygonal domain of $\mathbb{R}^2$. Let $\mathcal{T} = (\overline{\mathfrak{M}}, \overline{\mathfrak{M}^*}, \mathfrak{D})$ be a DDFV mesh satisfying* (5.5.1) *and* (5.5.2)*.*
Then for $p \geq 1$ and $q \geq 1$, there exists a constant $C > 0$ only depending on p, q, θ and Ω such that

$$\| u \|_{0,m,\mathcal{T}} \leq \frac{C}{(\sin(\alpha_{\mathcal{T}}))^{\theta/p} \zeta^{\theta(p-1)/(p)}} \| u \|_{1,p,\mathcal{T}}^{\theta} \| u \|_{0,q,\mathcal{T}}^{1-\theta}, \ \forall u \in Z(\mathcal{T}), \tag{5.5.3}$$

where

$$0 \leq \theta \leq \frac{p}{p + q(p-1)} \leq 1$$

and

$$\frac{1}{m} = \frac{1-\theta}{q} + \frac{\theta}{p} - \frac{\theta}{2}.$$

Proof. We start as in the proof of Theorem 5.3.1. Let $s \geq 1$. For $u \in Z(\mathcal{T})$, as $u_{\mathfrak{M}} \in X(\mathfrak{M})$ and $u_{\overline{\mathfrak{M}^*}} \in X(\overline{\mathfrak{M}^*})$, we may write:

$$\|u_{\mathfrak{M}}\|_{0,2s,\mathfrak{M}}^s \leq c(\Omega) \left(\left| |u_{\mathfrak{M}}|^s \right|_{1,1,\mathfrak{M}} + \|u_{\mathfrak{M}}\|_{0,s,\mathfrak{M}}^s \right) \tag{5.5.4}$$

$$\|u_{\overline{\mathfrak{M}^*}}\|_{0,2s,\overline{\mathfrak{M}^*}}^s \leq c(\Omega) \left(\left| |u_{\overline{\mathfrak{M}^*}}|^s \right|_{1,1,\overline{\mathfrak{M}^*}} + \|u_{\overline{\mathfrak{M}^*}}\|_{0,s,\overline{\mathfrak{M}^*}}^s \right) \tag{5.5.5}$$

But, following the same computations as in the proof of Theorem 5.3.1, we get

$$\begin{aligned} \left| |u_{\mathfrak{M}}|^s \right|_{1,1,\mathfrak{M}} &= \sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}_{int}} \mathrm{m}_\sigma \left| |u_{\mathcal{K}}|^s - |u_{\mathcal{L}}|^s \right| \\ &\leq s \sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}_{int}} \mathrm{m}_\sigma \mathrm{m}_{\sigma^*} \left| \frac{u_{\mathcal{K}} - u_{\mathcal{L}}}{\mathrm{m}_{\sigma^*}} \right| (|u_{\mathcal{K}}|^{s-1} + |u_{\mathcal{L}}|^{s-1}) \\ &\leq s \left(\sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}_{int}} \mathrm{m}_\sigma \mathrm{m}_{\sigma^*} \left| \frac{u_{\mathcal{K}} - u_{\mathcal{L}}}{\mathrm{m}_{\sigma^*}} \right|^p \right)^{\frac{1}{p}} \left(\sum_{\mathcal{K} \in \mathfrak{M}} \sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}_{\mathcal{K}}} \mathrm{m}_\sigma \mathrm{m}_{\sigma^*} |u_{\mathcal{K}}|^{\frac{(s-1)p}{p-1}} \right)^{\frac{p-1}{p}} \end{aligned}$$

Using the regularity hypotheses on the mesh, we get

$$\left||u_{\mathfrak{M}}|^s\right|_{1,1,\mathfrak{M}} \leq \frac{C}{(\sin(\alpha_{\mathcal{T}}))^{1/p}\zeta^{(p-1)/p}} \left(\sum_{\mathcal{D}_{\sigma,\sigma^*}\in\mathfrak{D}_{int}} \mathrm{m}_{\mathcal{D}} \left| \frac{u_{\mathcal{K}}-u_{\mathcal{L}}}{\mathrm{m}_{\sigma^*}} \right|^p \right)^{\frac{1}{p}} \|u_{\mathfrak{M}}\|^{s-1}_{0,(s-1)p/(p-1),\mathfrak{M}}.$$

But, by definition, $\dfrac{u_{\mathcal{K}}-u_{\mathcal{L}}}{\mathrm{m}_{\sigma^*}} = \nabla^{\mathcal{D}} u \cdot \boldsymbol{\tau}_{\mathcal{K},\mathcal{L}}$ and therefore $\left|\dfrac{u_{\mathcal{K}}-u_{\mathcal{L}}}{\mathrm{m}_{\sigma^*}}\right| \leq |\nabla^{\mathcal{D}} u|$. It yields :

$$\left||u_{\mathfrak{M}}|^s\right|_{1,1,\mathfrak{M}} \leq \frac{C}{(\sin(\alpha_{\mathcal{T}}))^{1/p}\zeta^{(p-1)/p}} |u|_{1,p,\mathcal{T}} \|u_{\mathfrak{M}}\|^{s-1}_{0,(s-1)p/(p-1),\mathfrak{M}}.$$

Let $q \geq 1$. We choose $s = 1 + (p-1)q/p \geq 1$. Injecting the last inequality in (5.5.4), and using the interpolation inequality (5.3.5), we get:

$$\begin{aligned} \|u_{\mathfrak{M}}\|^s_{0,2s,\mathfrak{M}} &\leq \frac{C}{(\sin(\alpha_{\mathcal{T}}))^{1/p}\zeta^{(p-1)/p}} \|u_{\mathfrak{M}}\|^{s-1}_{0,q,\mathfrak{M}} \left(|u|_{1,p,\mathcal{T}} + \|u_{\mathfrak{M}}\|_{0,p,\mathfrak{M}}\right) \\ &\leq \frac{C}{(\sin(\alpha_{\mathcal{T}}))^{1/p}\zeta^{(p-1)/p}} \|u_{\mathfrak{M}}\|^{s-1}_{0,q,\mathfrak{M}} \|u\|_{1,p,\mathcal{T}} \end{aligned}$$

because $\|u_{\mathfrak{M}}\|_{0,p,\mathfrak{M}} \leq 2\|u\|_{1,p,\mathcal{T}}$ by definition, with $p \geq 1$, $q \geq 1$ and $s = 1 + (p-1)q/p$. Then, let $0 \leq \alpha \leq 1$ and $m \geq 1$ such that

$$\frac{1}{m} = \frac{1-\alpha}{q} + \frac{\alpha}{2s}.$$

Using the interpolation inequality as in the proof of Theorem 5.3.1, we obtain:

$$\|u_{\mathfrak{M}}\|_{0,m,\mathfrak{M}} \leq \frac{C}{(\sin(\alpha_{\mathcal{T}}))^{\alpha/(sp)}\zeta^{\alpha(p-1)/(sp)}} \|u\|^{\alpha/s}_{1,p,\mathcal{T}} \|u\|^{1-\alpha/s}_{0,q,\mathcal{T}}, \quad \forall p \geq 1, \quad \forall q \geq 1,$$

with

$$s = 1 + \frac{(p-1)q}{p}, \quad \frac{1}{m} = \frac{1-\alpha}{q} + \frac{\alpha}{2s}.$$

With similar computations on the dual mesh, from (5.5.5), we get

$$\|u_{\overline{\mathfrak{M}^*}}\|_{0,m,\overline{\mathfrak{M}^*}} \leq \frac{C}{(\sin(\alpha_{\mathcal{T}}))^{\alpha/(sp)}\zeta^{\alpha(p-1)/(sp)}} \|u\|^{\alpha/s}_{1,p,\mathcal{T}} \|u\|^{1-\alpha/s}_{0,q,\mathcal{T}}.$$

Finally, setting $\theta = \alpha/s$, it yields the expected inequality (5.5.3). □

As in the classical finite volume framework, we can now prove discrete Sobolev-Poincaré inequalities. The proof is similar to the proof of Theorem 5.3.2; it will not be detailed here.

Theorem 5.5.2 (General discrete Sobolev-Poincaré inequality in the DDFV framework)**.** *Let Ω be an open bounded polygonal domain of $\mathbb{R}^2$. Let $\mathcal{T} = (\overline{\mathfrak{M}}, \overline{\mathfrak{M}^*}, \mathfrak{D})$ be a DDFV mesh satisfying* (5.5.1) *and* (5.5.2)*.*
Then there exists a constant $C > 0$ only depending on p, q and Ω such that:

– *if* $1 \leq p < 2$, *for all* $1 \leq q \leq \dfrac{2p}{2-p}$,

$$\| u \|_{0,q,\mathcal{T}} \leq \frac{C}{(\sin(\alpha_\mathcal{T}))^{1/p} \zeta^{(p-1)/(p)}} \| u \|_{1,p,\mathcal{T}}, \quad \forall u \in Z(\mathcal{T}),$$

– *if* $p \geq 2$, *for all* $1 \leq q < +\infty$,

$$\| u \|_{0,q,\mathcal{T}} \leq \frac{C}{(\sin(\alpha_\mathcal{T}))^{1/p} \zeta^{(p-1)/(p)}} \| u \|_{1,p,\mathcal{T}}, \quad \forall u \in Z(\mathcal{T}).$$

Let us now focus on the Poincaré-Wirtinger inequality in the DDFV case. This result has been proved recently in [133]. We will give here a proof using the embedding of $BV(\Omega)$ into $L^2(\Omega)$ (5.2.7) recalled in Theorem 5.2.2.

Theorem 5.5.3 (Discrete Poincaré-Wirtinger inequality in the DDFV framework)**.** *Let* Ω *be an open bounded connected polygonal domain of* $\mathbb{R}^2$. *Let* $\mathcal{T} = (\overline{\mathfrak{M}}, \overline{\mathfrak{M}^*}, \mathfrak{D})$ *be a DDFV mesh satisfying* (5.5.1).
There exists a constant $C > 0$ *depending only on* Ω, *such that for all* $u \in Z(\mathcal{T})$ *satisfying*

$$\sum_{\mathcal{K} \in \mathfrak{M}} \mathrm{m}_\mathcal{K} u_\mathcal{K} = \sum_{\mathcal{K} \in \overline{\mathfrak{M}^*}} \mathrm{m}_{\mathcal{K}^*} u_{\mathcal{K}^*} = 0, \tag{5.5.6}$$

we have

$$\|u\|_{0,2,\mathcal{T}} \leq \frac{C}{\sin(\alpha_\mathcal{T})} |u|_{1,2,\mathcal{T}}.$$

Proof. Let $u \in Z(\mathcal{T})$. Applying (5.2.7) to $u_\mathfrak{M} \in X(\mathfrak{M})$ and $u_{\overline{\mathfrak{M}^*}} \in X(\overline{\mathfrak{M}^*})$, we get, under the hypothesis (5.5.6),

$$\begin{aligned} \|u_\mathfrak{M}\|_{0,2,\mathfrak{M}} &\leq c(\Omega) TV_\Omega(u_\mathfrak{M}) &&\leq c(\Omega) \, |u_\mathfrak{M}|_{1,1,\mathfrak{M}} \\ \|u_{\overline{\mathfrak{M}^*}}\|_{0,2,\overline{\mathfrak{M}^*}} &\leq c(\Omega) TV_\Omega(u_{\overline{\mathfrak{M}^*}}) &&\leq c(\Omega) \, |u_{\overline{\mathfrak{M}^*}}|_{1,1,\overline{\mathfrak{M}^*}} \end{aligned}$$

But,

$$\begin{aligned} |u_\mathfrak{M}|_{1,1,\mathfrak{M}} &\leq \sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}_{int}} \mathrm{m}_\sigma \mathrm{m}_{\sigma^*} \frac{|u_\mathcal{K} - u_\mathcal{L}|}{\mathrm{m}_{\sigma^*}} \\ &\leq \frac{2}{\sin(\alpha_\mathcal{T})} \sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}_{int}} \mathrm{m}_\mathcal{D} \frac{|u_\mathcal{K} - u_\mathcal{L}|}{\mathrm{m}_{\sigma^*}} \\ &\leq \frac{2}{\sin(\alpha_\mathcal{T})} \mathrm{m}(\Omega)^{1/2} \, |u|_{1,2,\mathcal{T}}, \end{aligned}$$

thanks to Cauchy-Schwarz inequality. By the same way, we get the same bound for $|u_{\overline{\mathfrak{M}^*}}|_{1,1,\overline{\mathfrak{M}^*}}$ and it finally yields $\|u\|_{0,2,\mathcal{T}} \leq \dfrac{2}{\sin(\alpha_\mathcal{T})} \mathrm{m}(\Omega)^{1/2} c(\Omega) \, |u|_{1,2,\mathcal{T}}$. □

5.5.3 Discrete functional inequalities in the case with Dirichlet boundary conditions

In this section, we want to extend the discrete Gagliardo-Nirenberg-Sobolev inequalities of Section 5.4.2 to finite volume approximations obtained from a DDFV scheme. We first recall how Dirichlet boundary conditions are taken into account in DDFV methods. Let Γ_0 be a non empty part of the boundary. At the discrete level, homogeneous Dirichlet boundary conditions on Γ_0 will be written:

$$u_{\mathcal{K}} = 0,\ \forall \mathcal{K} \in \partial\mathfrak{M},\ \mathcal{K} \subset \Gamma^0 \text{ and } u_{\mathcal{K}^*} = 0,\ \forall \mathcal{K}^* \in \partial\mathfrak{M}^*,\ \overline{\mathcal{K}^*} \cap \Gamma^0 \neq \emptyset. \tag{5.5.7}$$

Therefore,we consider the corresponding set of finite volume approximations, $Z_{\Gamma^0}(\mathcal{T})$ defined by:

$$Z_{\Gamma^0}(\mathcal{T}) = \{u \in Z(\mathcal{T}) \text{ satisfying } (5.5.7)\}.$$

Let us note that the definition of the discrete $W^{1,p}$-seminorm is the same on $Z_{\Gamma^0}(\mathcal{T})$ as on $Z(\mathcal{T})$. Indeed, the fact that the approximate solution vanishes at the boundary is taken into account in the definition of the discrete gradient $\nabla^{\mathcal{D}} u$ for $\mathcal{D} \in \mathfrak{D}_{ext}$, and therefore in $|u|_{1,p,\mathcal{T}}$.

Finally, combining the techniques of proof of Theorem 5.4.1 (using Lemma 5.4.1) and Theorem 5.5.1, we establish the following theorem.

Theorem 5.5.4 (Discrete Gagliardo-Nirenberg-Sobolev inequality in the DDFV framework)**.** *Let Ω be an open convex bounded polygonal domain of $\mathbb{R}^2$ and Γ^0 be a part of the boundary. Let $\mathcal{T} = (\overline{\mathfrak{M}}, \overline{\mathfrak{M}^*}, \mathfrak{D})$ be a DDFV mesh satisfying* (5.5.1) *and* (5.5.2)*.*
Then for $p \geq 1$ and $q \geq 1$, there exists a constant $C > 0$ only depending on p, q and Ω such that

$$\| u \|_{0,m,\mathcal{T}} \leq \frac{C}{(\sin(\alpha_{\mathcal{T}}))^{\theta/p} \zeta^{\theta(p-1)/(p)}} \, | u |^{\theta}_{1,p,\mathcal{T}} \, \| u \|^{1-\theta}_{0,q,\mathcal{T}},\ \forall u \in Z_{\Gamma^0}(\mathcal{T}),$$

where

$$0 \leq \theta \leq \frac{p}{p + q(p-1)} \leq 1$$

and

$$\frac{1}{m} = \frac{1-\theta}{q} + \frac{\theta}{p} - \frac{\theta}{2}.$$

Chapitre 6

Un schéma volumes finis pour un modèle de Patlak-Keller-Segel avec diffusion croisée *

Nous analysons dans ce chapitre un schéma volumes finis pour un modèle de Keller-Segel parabolique-elliptique en dimension deux avec un terme de diffusion croisée additionnel dans l'équation elliptique sur le chimioattractant. L'ajout de ce terme de diffusion croisée empêche l'explosion des solutions en temps fini, qui peut avoir lieu pour le système de Keller-Segel classique. Le schéma proposé préserve la positivité, conserve la masse, et —sous des hypothèses additionnelles— dissipe l'entropie. Après avoir prouvé l'existence d'une solution au schéma par un argument de point fixe, nous obtenons une inégalité d'entropie–dissipation en utilisant des versions discrètes d'inégalités de Sobolev démontrées dans le chapitre 5, et nous pouvons finalement en déduire des estimations a priori qui permettent d'obtenir la compacité d'une famille de solutions approchées et de passer à la limite dans le schéma. Si de plus le paramètre de diffusion croisée est suffisamment important, nous prouvons la convergence en temps long de la solution approchée vers l'état stationnaire homogène en utilisant l'inégalité d'entropie. Enfin, nous présentons des simulations numériques qui semblent indiquer l'existence de solutions stationnaires non homogènes pour des valeurs intermédiaires du terme de diffusion croisée.

∗. Ce chapitre reprend des résultats obtenus en collaboration avec A. Jüngel. L'article correspondant, *A finite volume scheme for a Keller-Segel model with additional cross-diffusion* [20], est actuellement en cours de rédaction.

Contents

6.1 Introduction

Chemotaxis, the directed movement of cells in response to chemical gradients, plays an important role in many biological fields, such as embryogenesis, immunology, cancer growth, and wound healing [110, 151]. At the macroscopic level, chemotaxis models can be formulated in terms of the cell density $n(x,t)$ and the concentration of the chemical signal $S(x,t)$. A classical model to describe the time evolution of these two variables is the (Patlak-) Keller-Segel system, suggested by C. S. Patlak in 1953 [149] and E. F. Keller and L. A. Segel in 1970 [127]. Assuming that the time scale of the chemical signal is much larger than that of the cell movement, the classical parabolic-elliptic Keller-Segel equations read as follows:

$$\partial_t n = \operatorname{div}(\nabla n - n\nabla S), \quad 0 = \Delta S + \mu\, n - S \quad \text{in } \Omega,$$

where $\Omega \subset \mathbb{R}^2$ is a bounded domain or $\Omega = \mathbb{R}^2$. The parameter $\mu > 0$ is the secretion rate at which the chemical substance is emitted by the cells. The nonlinear term $n\nabla S$ models the cell movement towards higher concentrations of the chemical signal.

This model exhibits the phenomenon of cell aggregation. The more cells are aggregated, the more the attracting chemical signal is produced by the cells. This process is counterbalanced by cell diffusion, but if the cell density is sufficiently large, the nonlocal chemical interaction dominates diffusion and results in a blow-up of the cell density. In two space dimensions, the critical threshold for blow-up is given by $M = \int_\Omega n_0(x)\,dx = 4\pi$ if Ω is a bounded connected domain with C^2 boundary [142] and $M = 8\pi$ in the radial and whole-space case [22, 143]. The existence and uniqueness of smooth global-in-time

solutions in the subcritical case is proved for bounded domains in [116] and in the whole space in [24]. In the critical case $M = 8\pi$, a global whole-space solution exists, which becomes unbounded as $t \to \infty$ [23]. Furthermore, there exist radially symmetric initial data such that, in the supercritical case, the solution forms a δ-singularity in finite time [109].

Motivated by numerical and modeling issues, the question how blow up can be avoided has been investigated intensively the last years. It has been suggested to modify the chemotactic sensitivity (modeling, e.g., volume-filling effects), to allow for degenerate cell diffusion, or to include suitable growth-death terms. We refer to [111] for references. Another idea is to introduce additional cell diffusion in the equation for the chemical concentration [44, 111]. This diffusion term avoids, even for arbitrarily small diffusion constants, the blow-up and leads to the global-in-time existence of weak solutions [111]. The model, which is investigated in this chapter, reads as follows:

$$\partial_t n = \operatorname{div}(\nabla n - n\nabla S), \quad 0 = \Delta S + \delta\Delta n + \mu\, n - S, \quad x \in \Omega,\ t > 0, \tag{6.1.1}$$

where $\delta > 0$ is the additional diffusion constant. As usual, we impose homogeneous Neumann boundary and initial conditions

$$\nabla n \cdot \nu = \nabla S \cdot \nu = 0 \quad \text{on } \partial\Omega,\ t > 0, \quad n(\cdot, 0) = n_0 \quad \text{in } \Omega. \tag{6.1.2}$$

The advantage of the additional diffusion term is that blow-up of solutions is translated to large gradients which may help to determine the blow-up time numerically. Another advantage is that the enlarged system (6.1.1) exhibits an interesting entropy structure (see below).

At first sight, the additional term $\delta\Delta n$ seems to complicate the mathematical analysis. Indeed, the resulting diffusion matrix of the system is neither symmetric nor positive definite, and we cannot apply the maximum principle to the equation for the chemical signal anymore. It was shown in [111] that all these difficulties can be resolved by the observation that the above system possesses a logarithmic entropy,

$$E(t) = \int_\Omega \left(n\,(\log n - 1) + 1\right) dx,$$

which is dissipated according to

$$\frac{dE}{dt} + \int_\Omega \left(4\,|\nabla\sqrt{n}|^2 + \frac{1}{\delta}\,|\nabla S|^2 + \frac{1}{\delta}\,S^2\right) dx = \frac{\mu}{\delta}\int_\Omega n\,S\,dx. \tag{6.1.3}$$

Suitable Gagliardo-Nirenberg inequalities applied to the right-hand side lead to gradient estimates for $\sqrt{n}$ and S, which are the starting point for the existence and long-time analysis.

In this chapter, we aim at developing a finite volume scheme which preserves the entropy structure on the discrete level by generalizing the scheme proposed in [89]. In contrast to [89], we are able to prove the existence of discrete solutions and their numerical convergence to the continuous solution for all values of the initial mass. Moreover, we show

that the discrete solution converges for large times to the homogeneous steady state if μ or $1/\delta$ are sufficiently small.

In the literature, there exist several approaches to solve the classical Keller-Segel system numerically. The parabolic-elliptic model was approximated by using finite difference [157, 161] or finite element methods [139, 155, 159]. Also a dynamic moving-mesh method [35], a variational steepest descent approximation scheme [21], and a stochastic particle approximation [101, 102] were developed. Concerning numerical schemes for the parabolic-parabolic model (in which $\partial_t S$ is added to the second equation in (6.1.1)), we mention the second-order central-upwind finite volume method of [57], the discontinuous finite element approach of [78], and the conservative finite element scheme of [156]. We also cite the paper [36] for a mixed finite element discretization of a Keller-Segel model with nonlinear diffusion.

There are only a few works in which a numerical analysis of the scheme was performed. F. Filbet proved the existence and numerical convergence of finite volume solutions [89]. Error estimates for a conservative finite element approximation were shown by N. Saito [155, 156]. Y. Epshteyn and A. Izmirlioglu proved error estimates for a fully discrete discontinuous finite element method [78]. Convergence proofs for other schemes can be found in, e.g., [21, 102].

This chapter contains the first numerical analysis for the Keller-Segel model (6.1.1) with additional cross-diffusion. The originality comes from the fact that we "translate" all the analytical properties of [111] on a discrete level, namely positivity preservation, mass conservation, entropy stability, and entropy dissipation (under additional hypotheses).

The chapter is organized as follows. Section 6.2 is devoted to the description of the finite volume scheme and the formulation of the main results. The existence of a discrete solution is shown in Section 6.3. A discrete version of the entropy-dissipation relation (6.1.3) and corresponding gradient estimates are proved in Section 6.4. These estimates allow us to obtain in Section 6.5 the convergence of the discrete solution to the continuous one when the approximation parameters tend to zero. The long-time behavior of the discrete solution is investigated in Section 6.6. Finally, we present some numerical examples in Section 6.7 and compare the discrete solutions to our model (6.1.1) with those computed from the classical Keller-Segel system.

6.2 Numerical scheme and main results

In this section, we introduce the finite volume scheme and present our main results.

6.2.1 Notations and assumptions

Let $\Omega \subset \mathbb{R}^2$ be an open, bounded, polygonal subset. An admissible mesh of Ω is given by a family $\mathcal{T}$ of control volumes (open and convex polygons), a family $\mathcal{E}$ of edges, and a family of points $(x_K)_{K\in\mathcal{T}}$ which satisfy Definition 9.1 in [85]. This definition implies that the straight line between two neighboring centers of cells (x_K, x_L) is orthogonal to the edge $\sigma = K|L$. For instance, Voronoi meshes are admissible meshes [85, Example

9.2]. Triangular meshes satisfy the admissibility condition if all angles of the triangles are smaller than $\pi/2$ [85, Example 9.1].

We distinguish the interior edges $\sigma \in \mathcal{E}_{\rm int}$ and the boundary edges $\sigma \in \mathcal{E}_{\rm ext}$. The set of edges $\mathcal{E}$ equals the union $\mathcal{E}_{\rm int} \cup \mathcal{E}_{\rm ext}$. For a control volume $K \in \mathcal{T}$, we denote by $\mathcal{E}_K$ the set of its edges, by $\mathcal{E}_{{\rm int},K}$ the set of its interior edges, and by $\mathcal{E}_{{\rm ext},K}$ the set of edges of K included in $\partial\Omega$.

Furthermore, we denote by d the distance in $\mathbb{R}^2$ and by m the Lebesgue measure in $\mathbb{R}^2$ or $\mathbb{R}$. We assume that the family of meshes satisfies the following regularity requirement: there exists $\xi > 0$ such that for all $K \in \mathcal{T}$ and all $\sigma \in \mathcal{E}_{{\rm int},K}$ with $\sigma = K|L$, it holds

$$\mathrm{d}(x_K, \sigma) \geq \xi\, \mathrm{d}(x_K, x_L). \tag{6.2.1}$$

This hypothesis is needed to apply discrete Sobolev-type inequalities proved in Chapter 5. Introducing for $\sigma \in \mathcal{E}$ the notation

$$d_\sigma = \begin{cases} \mathrm{d}(x_K, x_L) & \text{if } \sigma \in \mathcal{E}_{\rm int},\ \sigma = K|L, \\ \mathrm{d}(x_K, \sigma) & \text{if } \sigma \in \mathcal{E}_{{\rm ext},K}, \end{cases}$$

we define the transmissibility coefficient

$$\tau_\sigma = \frac{\mathrm{m}(\sigma)}{d_\sigma}, \quad \sigma \in \mathcal{E}.$$

The size of the mesh is denoted by

$$\Delta x = \max_{K\in\mathcal{T}} \mathrm{diam}(K).$$

Let $T > 0$ be some final time and M_T the number of time steps. Then the time step size and the time points are given by, respectively,

$$\Delta t = \frac{T}{M_T}, \quad t^k = k\,\Delta t, \quad 0 \leq k \leq M_T.$$

We denote by $\mathcal{D}$ an admissible space-time discretization of $\Omega_T = \Omega \times (0,T)$ composed of an admissible mesh $\mathcal{T}$ of Ω and the values Δt and M_T. The size of this space-time discretization $\mathcal{D}$ is defined by $\eta = \max\{\Delta x, \Delta t\}$.

Let $X(\mathcal{T})$ be the linear space of functions $\Omega \to \mathbb{R}$ which are constant on each cell $K \in \mathcal{T}$. We define on $X(\mathcal{T})$ the discrete L^p norm, discrete $W^{1,p}$ seminorm, and discrete $W^{1,p}$ norm by, respectively,

$$\| u \|_{0,p,\mathcal{T}} = \left(\sum_{K\in\mathcal{T}} \mathrm{m}(K)\, |u|^p \right)^{1/p},$$

$$| u |_{1,p,\mathcal{T}} = \left(\sum_{\sigma\in\mathcal{E}_{\rm int}} \frac{\mathrm{m}(\sigma)}{d_\sigma^{p-1}}\, |D_\sigma u|^p \right)^{1/p},$$

$$\| u \|_{1,p,\mathcal{T}} = \| u \|_{0,p,\mathcal{T}} + | u |_{1,p,\mathcal{T}},$$

where $u \in X(\mathcal{T})$, $1 \leq p < \infty$, and $D_\sigma u = |u_K - u_L|$ for $\sigma = K|L \in \mathcal{E}_{\rm int}$.

6.2.2 Finite volume scheme and main results

We are now in the position to define the finite volume discretization of (6.1.1)-(6.1.2). Let $\mathcal{D}$ be a finite volume discretization of Ω_T. The initial datum n_0 is approximated by its L^2 projection on control volumes:

$$n_{\mathcal{D}}^0 = \sum_{K\in\mathcal{T}} n_K^0\,\mathbf{1}_K, \quad \text{where } n_K^0 = \frac{1}{\mathrm{m}(K)}\int_K n_0(x)\,dx, \tag{6.2.2}$$

and $\mathbf{1}_K$ is the characteristic function on K. Denoting by n_K^k and S_K^k approximations of the mean value of $n(\cdot,t^k)$ and $S(\cdot,t^k)$ on K, respectively, the numerical scheme reads as follows:

$$\mathrm{m}(K)\,\frac{n_K^{k+1}-n_K^k}{\Delta t} - \sum_{\sigma\in\mathcal{E}_K}\tau_\sigma\,Dn_{K,\sigma}^{k+1} + \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}}\tau_\sigma\left((DS_{K,\sigma}^{k+1})^+ n_K^{k+1} - (DS_{K,\sigma}^{k+1})^- n_L^{k+1}\right) = 0, \tag{6.2.3}$$

$$-\sum_{\sigma\in\mathcal{E}_K}\tau_\sigma\,DS_{K,\sigma}^{k+1} - \delta\sum_{\sigma\in\mathcal{E}_K}\tau_\sigma\,Dn_{K,\sigma}^{k+1} = \mathrm{m}(K)\left(\mu\,n_K^{k+1} - S_K^{k+1}\right), \tag{6.2.4}$$

for all $K\in\mathcal{T}$ and $0\le k\le M_T-1$. Here, $v^+=\max\{0,v\}$, $v^-=\max\{0,-v\}$, and

$$DU_{K,\sigma}^k = \begin{cases} U_L^k - U_K^k & \text{for } \sigma = K|L\in\mathcal{E}_{{\rm int},K},\\ 0 & \text{for } \sigma\in\mathcal{E}_{{\rm ext},K}.\end{cases} \tag{6.2.5}$$

The approximation S_K^0 is computed from (6.2.4) with $k=-1$. This scheme is based on a fully implicit Euler discretization in time and a finite volume approach for the volume variable. The implicit scheme allows us to establish discrete entropy-dissipation estimates which would not be possible with an explicit scheme. This approximation is similar to that in [89] except the additional cross-diffusion term in the second equation.

The numerical approximations $n_{\mathcal{D}}$ and $S_{\mathcal{D}}$ of n and S are defined by

$$n_{\mathcal{D}}(x,t) = \sum_{K\in\mathcal{T}} n_K^{k+1}\,\mathbf{1}_K(x), \quad S_{\mathcal{D}}(x,t) = \sum_{K\in\mathcal{T}} S_K^{k+1}\,\mathbf{1}_K(x), \quad \text{where } x\in\Omega,\ t\in(t^k,t^{k+1}),$$

and $k=0,\dots,M_T-1$. Furthermore, we define approximations $\nabla^{\mathcal{D}}n_{\mathcal{D}}$ and $\nabla^{\mathcal{D}}S_{\mathcal{D}}$ of the gradients of n and S, respectively. To this end, we introduce a dual mesh: for $K\in\mathcal{T}$ and $\sigma\in\mathcal{E}_K$, let $T_{K,\sigma}$ be defined by:

- If $\sigma = K|L\in\mathcal{E}_{{\rm int},K}$, $T_{K,\sigma}$ is the cell ("diamond") whose vertices are given by x_K, x_L, and the end points of the edge $\sigma = K|L$.
- If $\sigma\in\mathcal{E}_{{\rm ext},K}$, $T_{K,\sigma}$ is the cell ("triangle") whose vertices are given by x_K and the end points of the edge $\sigma = K|L$.

An example of construction of $T_{K,\sigma}$ can be found in [52]. Clearly, $T_{K,\sigma}$ defines a partition of Ω. The approximate gradient $\nabla^{\mathcal{D}}n_{\mathcal{D}}$ is a piecewise constant function, defined in $\Omega_T = \Omega\times(0,T)$ by

$$\nabla^{\mathcal{D}}n_{\mathcal{D}}(x,t) = \frac{\mathrm{m}(\sigma)}{\mathrm{m}(T_{K,\sigma})}\,Dn_{K,\sigma}^{k+1}\,\nu_{K,\sigma}, \quad x\in T_{K,\sigma},\ t\in(t^k,t^{k+1}),$$

where $Dn_{K,\sigma}^{k+1}$ is given as in (6.2.5) and $\nu_{K,\sigma}$ is the unit vector normal to σ and outward to K. The approximate gradient $\nabla^{\mathcal{D}} S_{\mathcal{D}}$ is defined in a similar way.

Our first result is the existence of solutions to the finite volume scheme.

Theorem 6.2.1 (Existence of finite volume solutions)**.** *Let $\Omega \subset \mathbb{R}^2$ be an open, bounded, polygonal subset and let $\mathcal{D}$ be an admissible discretization of $\Omega \times (0,T)$. The initial datum satisfies $n_0 \in L^2(\Omega)$, $n_0 \geq 0$ in Ω. Then there exists a solution $\{(n_K^k, S_K^k),\, K \in \mathcal{T},\, 0 \leq k \leq M_T\}$ to* (6.2.2)-(6.2.4) *satisfying*

$$n_K^k \geq 0, \quad \text{for all } K \in \mathcal{T},\, k \geq 0, \tag{6.2.6}$$

$$\sum_{K\in\mathcal{T}} \mathrm{m}(K)\, n_K^k = \sum_{K\in\mathcal{T}} \mathrm{m}(K)\, n_K^0 = \|n_0\|_{L^1(\Omega)}, \quad \text{for all } k \geq 0. \tag{6.2.7}$$

Properties (6.2.6) and (6.2.7) show that the scheme is positivity preserving and mass conserving. It is also entropy stable; see (6.4.1) below.

Let $(\mathcal{D}_\eta)_{\eta>0}$ be a sequence of admissible space-time discretizations indexed by the size $\eta = \max\{\Delta x, \Delta t\}$ of the discretization. We denote by $(\mathcal{T}_\eta)_{\eta>0}$ the corresponding meshes of Ω. We suppose that these discretizations satisfy (6.2.1) uniformly in η, i.e., $\xi > 0$ does not depend on η. Let $(n_\eta, S_\eta) := (n_{\mathcal{D}_\eta}, S_{\mathcal{D}_\eta})$ be a finite volume solution, constructed in Theorem 6.2.1, on the discretization $\mathcal{D}_\eta$. We set $\nabla^\eta := \nabla^{\mathcal{D}_\eta}$. Our second result concerns the convergence of (n_η, S_η) to a weak solution (n, S) to (6.1.1)-(6.1.2).

Theorem 6.2.2 (Convergence of the finite volume solutions)**.** *Let the assumptions of Theorem 6.2.1 hold. Furthermore, let $(\mathcal{D}_\eta)_{\eta>0}$ be a sequence of admissible discretizations satisfying* (6.2.1) *uniformly in η, and let (n_η, S_η) be a sequence of finite volume solutions to* (6.2.2)-(6.2.4)*. Then there exists (n, S) such that, up to a subsequence,*

$$\begin{aligned} n_\eta &\to n \quad \text{strongly in } L^2(\Omega_T), \\ \nabla^\eta n_\eta &\rightharpoonup \nabla n \quad \text{weakly in } L^2(\Omega_T), \\ S_\eta \rightharpoonup S, \quad \nabla^\eta S_\eta &\rightharpoonup \nabla S \quad \text{weakly in } L^2(\Omega_T), \end{aligned}$$

and $(n, S) \in L^2(0,T; H^1(\Omega))^2$ is a weak solution to (6.1.1)-(6.1.2) *in the sense of*

$$\int_0^T \int_\Omega (n\, \partial_t \phi - \nabla n \cdot \nabla \phi + n \nabla S \cdot \nabla \phi)\, dx + \int_\Omega n_0\, \phi(\cdot, 0)\, dx = 0, \tag{6.2.8}$$

$$\int_0^T \int_\Omega (-\nabla S \cdot \nabla \phi - \delta\, \nabla n \cdot \nabla \phi + \mu\, n\, \phi - S\, \phi)\, dx = 0 \tag{6.2.9}$$

for all test functions $\phi \in C_0^\infty(\Omega \times [0,T])$.

It is shown in [111, Theorem 1.3] that, if the secretion rate $\mu > 0$ is sufficiently small or the diffusion parameter $\delta > 0$ is sufficiently large, the solution (n, S) to (6.1.1)-(6.1.2) converges exponentially fast to the homogeneous steady state (n^*, S^*), where $n^* = \|n_0\|_{L^1(\Omega)}/\mathrm{m}(\Omega)$ and $S^* = \mu\, n^*$. We prove a related result for the finite volume solutions, but without convergence rate. We have not been able to show exponential convergence

since the proof in [111] is based on the logarithmic Sobolev inequality and we do not dispose of a discrete version. We introduce the discrete relative entropy

$$E[n^k|n^*] = \sum_{K\in\mathcal{T}} \mathrm{m}(K)\, n_K^k \,\log\left(\frac{n_K^k}{n^*}\right) \geq 0, \quad k \geq 0.$$

Theorem 6.2.3 (Long-time behavior of finite volume solutions)**.** *Let the assumptions of Theorem 6.2.1 hold and let* $(n_\mathcal{D}, S_\mathcal{D})$ *be a solution to* (6.2.2)-(6.2.4)*. Then there exists a constant* $C(\Omega) > 0$ *only depending on* Ω *such that for all* $k \geq 0$*,*

$$E[n^{k+1}|n^*] + \Delta t\,(1 - C^*)\,\left|\sqrt{n^{k+1}}\right|_{1,2,\mathcal{T}}^2 + \frac{\Delta t}{2\delta}\,\left\|S^{k+1} - S^*\right\|_{1,2,\mathcal{T}}^2 \leq E[n^k|n^*],$$

where $C^* = 2\,\mu^2\,C(\Omega)^2\,\|n_0\|_{L^1(\Omega)}/(\delta\,\xi)$ *and* ξ *is the parameter in* (6.2.1)*. In particular, if* $C^* < 1$*, the discrete relative entropy is nonincreasing and for each* $K \in \mathcal{T}$*,*

$$(n_K^k, S_K^k) \to (n^*, S^*) \quad \text{as } k \to \infty.$$

In the above theorem, $C(\Omega) > 0$ is the constant in the following discrete Poincaré-Sobolev inequality (Theorem 5.3.2

$$\left\|n^{k+1} - n^*\right\|_{0,2,\mathcal{T}} \leq C(\Omega)\,\left|n^{k+1}\right|_{1,1,\mathcal{T}}.$$

In [111], the long-time behavior of solutions is shown under the condition $\mu^2\|n_0\|_{L^1(\Omega)}/\delta < C_1(\Omega)$ for a constant $C_1(\Omega) > 0$ appearing in some Poincaré-Sobolev inequality. We observe that our condition depends additionally on the regularity of the finite volume mesh.

6.3 Existence of finite volume solutions

In this section, we prove Theorem 6.2.1. The proof is based on the Brouwer fixed-point theorem. Let $k \in \{1, \ldots, M_T - 1\}$ and let (n^k, S^k) be a solution to (6.2.3) and (6.2.4), with $k+1$ replaced by k, satisfying (6.2.6)-(6.2.7). We introduce the set

$$Z = \{u \in X(\mathcal{T}) : u \geq 0 \text{ in } \Omega,\ \|u\|_{L^1(\Omega)} \leq \|n_0\|_{L^1(\Omega)}\}.$$

The finite-dimensional space Z is convex and compact. In the following, we define the fixed-point operator by solving a linearized problem. First, we construct $\widetilde{S} \in X(\mathcal{T})$ using the following scheme:

$$-\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\, D\widetilde{S}_{K,\sigma} + \mathrm{m}(K)\,\widetilde{S}_K = \delta \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\, Dn_{K,\sigma}^k + \mu\,\mathrm{m}(K)\,n_K^k, \quad K \in \mathcal{T}. \tag{6.3.1}$$

Second, we compute $\widetilde{n} \in X(\mathcal{T})$ using the scheme:

$$\frac{\mathrm{m}(K)}{\Delta t}\,(\widetilde{n}_K - n_K^k) - \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\, D\widetilde{n}_{K,\sigma} + \sum_{\substack{\sigma\in\mathcal{E}_{\mathrm{int}},\\ \sigma=K|L}} \tau_\sigma\,\left((D\widetilde{S}_{K,\sigma})^+\widetilde{n}_K - (D\widetilde{S}_{K,\sigma})^-\widetilde{n}_L\right) = 0. \tag{6.3.2}$$

Step 1: Existence and uniqueness for (6.3.1) *and* (6.3.2). The linear system (6.3.1) can be written as $A\,\widetilde{S} = b$, where A is the matrix with elements

$$A_{K,K} = \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma + \mathrm{m}(K), \quad A_{K,L} = -\tau_\sigma \quad \text{for } K, L \in \mathcal{T} \text{ with } \sigma = K|L \in \mathcal{E}_{\mathrm{int},K},$$

and b is the vector with elements

$$b_K = \delta \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\, Dn^k_{K,\sigma} + \mu\,\mathrm{m}(K)\, n^k_K.$$

Since for all $L \in \mathcal{T}$,

$$|A_{L,L}| - \sum_{K\neq L} |A_{K,L}| = \mathrm{m}(L) > 0,$$

the matrix A is strictly diagonally dominant with respect to the columns and hence, A is invertible. This shows the unique solvability of (6.3.1).

Similarly, (6.3.2) can be written as $B\,\widetilde{n} = c$, where B is the matrix with elements

$$B_{K,K} = \frac{\mathrm{m}(K)}{\Delta t} + \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma \left(1 + (D\widetilde{S}_{K,\sigma})^+\right), \quad K \in \mathcal{T},$$
$$B_{K,L} = -\tau_\sigma \left(1 + (D\widetilde{S}_{K,\sigma})^-\right), \quad \text{for } K \in \mathcal{T},\ L \in \mathcal{T} \text{ with } \sigma = K|L \in \mathcal{E}_{\mathrm{int},K},$$

and c is the vector with elements $c_K = \mathrm{m}(K)n^k_K/\Delta t$, $K \in \mathcal{T}$. The diagonal elements of B are positive, and the off-diagonal elements are nonpositive. Moreover, B is strictly diagonal dominant with respect to the columns since for all $\sigma = K|L \in \mathcal{E}_{\mathrm{int},K}$, we have $D\widetilde{S}_{L,\sigma} = -D\widetilde{S}_{K,\sigma}$ which yields $(D\widetilde{S}_{L,\sigma})^+ = (D\widetilde{S}_{K,\sigma})^-$ and hence,

$$|B_{L,L}| - \sum_{K\neq L} |B_{K,L}| = \frac{\mathrm{m}(L)}{\Delta t} > 0.$$

We infer that B is an M-matrix and invertible, which gives the existence and uniqueness of a solution $\widetilde{n}$ to (6.3.2).

The M-matrix property of B implies that B^{-1} is positive. As a consequence, since n^k and c are nonnegative componentwise, by the induction hypothesis, $\widetilde{n} = B^{-1}c$ is nonnegative componentwise. This means that $\widetilde{n}$ satisfies (6.2.6). Summing (6.3.2) over $K \in \mathcal{T}$, we compute

$$\sum_{K\in\mathcal{T}} \mathrm{m}(K)\,\widetilde{n}_K = \sum_{K\in\mathcal{T}} \mathrm{m}(K)\, n^k_K = \|n_0\|_{L^1(\Omega)}.$$

Step 2: Continuity of the fixed-point operator. The solution to (6.3.1) and (6.3.2) defines the fixed-point operator $F : Z \to Z$, $F(n) = \widetilde{n}$. We have to show that F is continuous. For this, let $(n^\gamma)_{\gamma\in\mathbb{N}} \subset Z$ be a sequence converging to n in $X(\mathcal{T})$ as $\gamma \to \infty$. Setting $\widetilde{n}^\gamma = F(n^\gamma)$ and $\widetilde{n} = F(n)$, we have to prove that $\widetilde{n}^\gamma \to \widetilde{n}$ in $X(\mathcal{T})$. Using the scheme (6.3.1), we construct first $\widetilde{S}^\gamma$ (respectively, $\widetilde{S}$) from n^γ (respectively, n). Then, using the scheme (6.3.2), we obtain $\widetilde{n}^\gamma$ (respectively, $\widetilde{n}$).

We claim that $\widetilde{S}^\gamma - \widetilde{S} \to 0$ in $X(\mathcal{T})$ as $\gamma \to \infty$. Indeed, since the map $n \mapsto \tilde{S}$, where $\tilde{S}$ is constructed from (6.3.1), is linear on the finite dimensional space $X(\mathcal{T})$, it is obviously continuous.

Moreover, using the scheme (6.3.2) and performing the same computations as in the proof of Theorem 2.1 in [89], it follows that

$$\sum_{K\in\mathcal{T}} \mathrm{m}(K)\ |\widetilde{n}_K^\gamma - \widetilde{n}_K| \le 2\,\Delta t \left(\sum_{K\in\mathcal{T}} |\widetilde{n}_K|^2\right)^{1/2} \left(\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma \left|D(\widetilde{S}^\gamma - \widetilde{S})_{K,\sigma}\right|^2\right)^{1/2}.$$

The right-hand side converges to zero as $\widetilde{S}^\gamma \to \widetilde{S}$ in $X(\mathcal{T})$, which proves that $\widetilde{n}^\gamma \to \widetilde{n}$ in $X(\mathcal{T})$.

Step 3: Application of the fixed-point theorem. The assumptions of the Brouwer fixed-point theorem are satisfied, implying the existence of a fixed point of F, i.e. of a solution n^{k+1} to (6.2.3) satisfying (6.2.6). We have shown in Step 1 that (6.2.7) holds for n^{k+1}. Finally, we construct S^{k+1} using scheme (6.2.4).

6.4 A priori estimates

The proof of Theorem 6.2.2 is based on suitable a priori estimates which are shown in this section. We introduce a discrete version of the entropy functional used in [111]:

$$E^k = \sum_{K\in\mathcal{T}} \mathrm{m}(K)\, H\left(n_K^k\right), \quad \text{where } H(s) = s\,(\log s - 1) + 1.$$

Proposition 6.4.1 (Entropy stability)**.** *There exists a constant $C > 0$ only depending on Ω, μ, δ, $\|n_0\|_{L^1(\Omega)}$, and ξ (see* (6.2.1)*) such that for all $k \ge 0$,*

$$E^{k+1} - E^k + \frac{\Delta t}{2} \sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma \left|D\left(\sqrt{n^{k+1}}\right)_{K,\sigma}\right|^2 + \frac{\Delta t}{\delta} \sum_{K\in\mathcal{T}} \mathrm{m}(K)\ \left|S^{k+1}\right|^2 + \frac{\Delta t}{\delta} \sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma \left|DS_{K,\sigma}^{k+1}\right|^2 \le C\,\Delta t. \tag{6.4.1}$$

Proof. By the convexity of H, we find that

$$E^{k+1} - E^k = \sum_{K\in\mathcal{T}} \mathrm{m}(K)\left(H\left(n_K^{k+1}\right) - H\left(n_K^k\right)\right) \le \sum_{K\in\mathcal{T}} \mathrm{m}(K) \log\left(n_K^{k+1}\right)\left(n_K^{k+1} - n_K^k\right).$$

Inserting the scheme (6.2.3), we can write

$$E^{k+1} - E^k \le \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma \left(n_L^{k+1} - n_K^{k+1}\right) \log n_K^{k+1} - \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma \left(\left(DS_{K,\sigma}^{k+1}\right)^+ n_K^{k+1} - \left(DS_{K,\sigma}^{k+1}\right)^- n_L^{k+1}\right) \log n_K^{k+1} =: I_1 + I_2. \tag{6.4.2}$$

Now, we argue similarly as in the proof of Lemma 3.1 in [89]. We employ the symmetry of τ_σ and a Taylor expansion of log around n_K^{k+1} to infer that

$$\begin{aligned} I_1 &= -\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma \left(n_K^{k+1}-n_L^{k+1}\right)\left(\log n_K^{k+1}-\log n_L^{k+1}\right) \\ &= -\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma \left(n_K^{k+1}-n_L^{k+1}\right)^2 \frac{1}{\bar n_\sigma^{k+1}} = -\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma \left(\frac{Dn_{K,\sigma}^{k+1}}{\sqrt{\bar n_\sigma^{k+1}}}\right)^2, \end{aligned}$$

where $\bar n_\sigma^{k+1} = t_\sigma\, n_K^{k+1} + (1-t_\sigma)\, n_L^{k+1}$ for some $t_\sigma\in(0,1)$. We perform a summation by parts in I_2, using again the symmetry of τ_σ:

$$I_2 = -\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma \left(\left(DS_{K,\sigma}^{k+1}\right)^+ n_K^{k+1} - \left(DS_{K,\sigma}^{k+1}\right)^- n_L^{k+1}\right)\left(\log n_K^{k+1}-\log n_L^{k+1}\right).$$

Reordering the sum and using the expression for $\bar n_\sigma^{k+1}$ in the Taylor expansion of log, it is shown in [89, p. 468] that

$$I_2 \le -\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma\, \bar n_\sigma^{k+1}\, DS_{K,\sigma}^{k+1}\left(\log n_K^{k+1}-\log n_L^{k+1}\right).$$

The Taylor expansion shows that $\bar n_\sigma^{k+1}\left(\log n_K^{k+1}-\log n_L^{k+1}\right) = n_K^{k+1}-n_L^{k+1}$, which gives

$$I_2 \le -\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma\, DS_{K,\sigma}^{k+1}\left(n_K^{k+1}-n_L^{k+1}\right) = \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma\, DS_{K,\sigma}^{k+1}\, Dn_{K,\sigma}^{k+1}.$$

Summarizing the estimates for I_1 and I_2, (6.4.2) leads to

$$E^{k+1}-E^k \le -\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma \left(\frac{Dn_{K,\sigma}^{k+1}}{\sqrt{\bar n_\sigma^{k+1}}}\right)^2 + \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma\, DS_{K,\sigma}^{k+1}\, Dn_{K,\sigma}^{k+1}. \tag{6.4.3}$$

The first term can be estimated for $\sigma=K|L$ as follows:

$$\frac{|Dn_{K,\sigma}^{k+1}|}{\sqrt{\bar n_\sigma^{k+1}}} = \frac{\sqrt{n_L^{k+1}}+\sqrt{n_K^{k+1}}}{\sqrt{\bar n_\sigma^{k+1}}}\left|D(\sqrt{n^{k+1}})_{K,\sigma}\right| \ge \left|D(\sqrt{n^{k+1}})_{K,\sigma}\right|. \tag{6.4.4}$$

In order to bound the second term, we multiply the scheme (6.2.4) by $(\Delta t/\delta)\, S_K^{k+1}$ and sum over $K\in\mathcal{T}$:

$$\begin{aligned} 0 &= \frac{\Delta t}{\delta}\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\, DS_{K,\sigma}^{k+1}\, S_K^{k+1} + \Delta t \sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\, Dn_{K,\sigma}^{k+1}\, S_K^{k+1} \\ &\quad + \frac{\mu\,\Delta t}{\delta}\sum_{K\in\mathcal{T}} \mathrm{m}(K)\, n_K^{k+1}\, S_K^{k+1} - \frac{\Delta t}{\delta}\sum_{K\in\mathcal{T}} \mathrm{m}(K)\left|S_K^{k+1}\right|^2. \end{aligned}$$

By summation by parts, we find that

$$\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma\, Dn_{K,\sigma}^{k+1}\, DS_{K,\sigma}^{k+1} = -\frac{\Delta t}{\delta} \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma \left|DS_{K,\sigma}^{k+1}\right|^2 + \frac{\mu\,\Delta t}{\delta} \sum_{K\in\mathcal{T}} \mathrm{m}(K)\, n_K^{k+1}\, S_K^{k+1}$$

$$- \frac{\Delta t}{\delta} \sum_{K\in\mathcal{T}} \mathrm{m}(K) \left|S_K^{k+1}\right|^2. \tag{6.4.5}$$

Inserting (6.4.5) into (6.4.3) and employing (6.4.4), it follows that

$$E^{k+1} - E^k + \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma \left|D\left(\sqrt{n^{k+1}}\right)_{K,\sigma}\right|^2 + \frac{\Delta t}{\delta} \sum_{K\in\mathcal{T}} \mathrm{m}(K) \left|S_K^{k+1}\right|^2$$

$$+ \frac{\Delta t}{\delta} \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma \left|DS_{K,\sigma}^{k+1}\right|^2 \le \frac{\mu\Delta t}{\delta} \sum_{K\in\mathcal{T}} \mathrm{m}(K) n_K^{k+1} S_K^{k+1}. \tag{6.4.6}$$

It remains to estimate the right-hand side. We follow the proof of Proposition 3.1 in [111]. The Hölder inequality yields

$$\mu \sum_{K\in\mathcal{T}} \mathrm{m}(K)\, n_K^{k+1}\, S_K^{k+1} \;\le\; \mu \left\|n^{k+1}\right\|_{0,6/5,\mathcal{T}} \left\|S^{k+1}\right\|_{0,6,\mathcal{T}}. \tag{6.4.7}$$

The discrete L^6 norm of S^{k+1} can be bounded by the discrete H^1 norm using the discrete Sobolev inequality (Theorem 5.3.2):

$$\left\|S^{k+1}\right\|_{0,6,\mathcal{T}} \le C \left\|S^{k+1}\right\|_{1,2,\mathcal{T}},$$

where $C > 0$ depends only on Ω and ξ. For the discrete $L^{6/5}$ norm of n^{k+1}, we employ the discrete Gagliardo-Nirenberg inequality (Theorem 5.3.1) with $\theta = 1/6$:

$$\left\|n^{k+1}\right\|_{0,6/5,\mathcal{T}} = \left\|\sqrt{n^{k+1}}\right\|_{0,12/5,\mathcal{T}}^2 \le C \left\|\sqrt{n^{k+1}}\right\|_{0,2,\mathcal{T}}^{2(1-\theta)} \left\|\sqrt{n^{k+1}}\right\|_{1,2,\mathcal{T}}^{2\theta},$$

where $C > 0$ depends only on Ω and ξ. Mass conservation (6.2.7) implies that

$$\left\|n^{k+1}\right\|_{0,6/5,\mathcal{T}} \le C \left\|n_0\right\|_{L^1(\Omega)}^{1-\theta} \left(\left\|n_0\right\|_{L^1(\Omega)}^{1/2} + \left|\sqrt{n^{k+1}}\right|_{1,2,\mathcal{T}}\right)^{2\theta}.$$

With these estimates, (6.4.7) becomes

$$\mu \sum_{K\in\mathcal{T}} \mathrm{m}(K)\, n_K^{k+1}\, S_K^{k+1} \le C\,\mu \left\|S^{k+1}\right\|_{1,2,\mathcal{T}} \left\|n_0\right\|_{L^1(\Omega)}^{1-\theta} \left(\left\|n_0\right\|_{L^1(\Omega)}^{1/2} + \left|\sqrt{n^{k+1}}\right|_{1,2,\mathcal{T}}\right)^{2\theta}$$

$$\le\; 2\,C\,\mu \left\|S^{k+1}\right\|_{1,2,\mathcal{T}} \left\|n_0\right\|_{L^1(\Omega)}^{1-\theta} \left(\left\|n_0\right\|_{L^1(\Omega)} + \left|\sqrt{n^{k+1}}\right|_{1,2,\mathcal{T}}^2\right)^{\theta}$$

Then, by Young's inequality with $p_1 = 2$, $p_2 = 2/(1-2\theta)$, $p_3 = 1/\theta$,

$$\begin{aligned}\mu \sum_{K\in\mathcal{T}} \mathrm{m}(K)\, n_K^{k+1}\, S_K^{k+1} &\le \frac{1}{2}\left\|S^{k+1}\right\|_{1,2,\mathcal{T}}^2 + C(\delta,\mu)\,\|n_0\|_{L^1(\Omega)}^{2(1-\theta)/(1-2\theta)} \\ &\quad + \frac{\delta}{2}\left(\|n_0\|_{L^1(\Omega)} + \left|\sqrt{n^{k+1}}\right|_{1,2,\mathcal{T}}^2\right) \\ &\le \frac{1}{2}\left\|S^{k+1}\right\|_{1,2,\mathcal{T}}^2 + \frac{\delta}{2}\left|\sqrt{n^{k+1}}\right|_{1,2,\mathcal{T}}^2 + C(\delta,\mu,\|n_0\|_{L^1(\Omega)}).\end{aligned}$$

Together with (6.4.6), this finishes the proof. □

Summing (6.4.1) over $k = 0,\ldots,M_T-1$ and using the mass conservation (6.2.7), we conclude immediately the following η-uniform bounds for the family of solutions (n_η, S_η) to (6.2.2)-(6.2.4) with discretizations $\mathcal{D}_\eta$:

$$\begin{gathered}(n_\eta),\ (n_\eta \log n_\eta) \text{ are bounded in } L^\infty(0,T;L^1(\Omega)), \\ (\nabla^\eta n_\eta) \text{ is bounded in } L^2(\Omega_T), \\ (S_\eta) \text{ is bounded in } L^2(0,T;H^1(\Omega)).\end{gathered} \tag{6.4.8}$$

Then we can also deduce the following η-uniform bound for (n_η):

Proposition 6.4.2. *The family $(n_\eta)_{\eta>0}$ is bounded in $L^2(0,T;H^1(\Omega))$.*

Proof. First, we claim that (n_η) is bounded in $L^2(0,T;W^{1,1}(\Omega))$. To simplify the notation, we write $n_\eta^{k+1} := n_\eta(\cdot,t^{k+1}) \in X(\mathcal{T}_\eta)$. Applying the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\left|n_\eta^{k+1}\right|_{1,1,\mathcal{T}_\eta} &= \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \mathrm{m}(\sigma)\,\left|n_L^{k+1} - n_K^{k+1}\right| \\ &\le \sum_{K\in\mathcal{T}_\eta}\sum_{\sigma\in\mathcal{E}_K} \sqrt{\tau_\sigma}\left|D\left(\sqrt{n^{k+1}}\right)_{K,\sigma}\right| \cdot \sqrt{\mathrm{m}(\sigma)\,d_\sigma}\,\sqrt{n_K^{k+1}} \\ &\le \left|\sqrt{n_\eta^{k+1}}\right|_{1,2,T_\eta}\left(\sum_{K\in\mathcal{T}_\eta}\left(\sum_{\sigma\in\mathcal{E}_K}\mathrm{m}(\sigma)\,d_\sigma\right) n_K^{k+1}\right)^{1/2}.\end{aligned}$$

Observe that in two space dimensions,

$$\sum_{\sigma\in\mathcal{E}_K} \mathrm{m}(\sigma)\,\mathrm{d}(x_K,\sigma) = 2\,\mathrm{m}(K),$$

since the straight line between x_K and x_L is orthogonal to the edge $\sigma = K|L$. Using this property, the mesh regularity assumption (6.2.1), and the mass conservation (6.2.7), it follows that

$$\begin{aligned}\left|n_\eta^{k+1}\right|_{1,1,\mathcal{T}_\eta} &\le \left(\frac{2}{\xi}\right)^{1/2}\left|\sqrt{n_\eta^{k+1}}\right|_{1,2,\mathcal{T}_\eta}\left(\sum_{K\in\mathcal{T}_\eta}\mathrm{m}(K)\,n_K^{k+1}\right)^{1/2} \\ &= \left(\frac{2}{\xi}\right)^{1/2}\left|\sqrt{n_\eta^{k+1}}\right|_{1,2,\mathcal{T}_\eta}\|n_0\|_{L^1(\Omega)}^{1/2}.\end{aligned} \tag{6.4.9}$$

In view of the entropy stability estimate (6.4.1), we infer that (n_η) is bounded in $L^2(0,T;W^{1,1}(\Omega))$. Because of the discrete Sobolev inequality (Theorem 5.3.2)

$$\left\|n_\eta^{k+1}\right\|_{0,2,\mathcal{T}_\eta} \le C\,\left\|n_\eta^{k+1}\right\|_{1,1,T_\eta},$$

the family (n_η) is bounded in $L^2(\Omega_T)$.

In order to estimate the approximate gradient of n_η, we employ (6.4.3). The last term in (6.4.3) is treated as follows. We multiply (6.2.4) by $\Delta t\, n_K^{k+1}$, sum over $K\in\mathcal{T}$, and sum by parts:

$$\begin{aligned}\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma\, DS_{K,\sigma}^{k+1}\, Dn_{K,\sigma}^{k+1} &= -\delta\,\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma \left|Dn_{K,\sigma}^{k+1}\right|^2 \\ &\quad + \mu\,\Delta t \sum_{K\in\mathcal{T}_\eta} \mathrm{m}(K)\left|n_K^{k+1}\right|^2 - \Delta t \sum_{K\in\mathcal{T}_\eta} \mathrm{m}(K)\, S_K^{k+1}\, n_K^{k+1} \\ &\le -\delta\,\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma \left|Dn_{K,\sigma}^{k+1}\right|^2 + C\left(\left\|n_\eta^{k+1}\right\|_{0,2,\mathcal{T}_\eta}^2 + \left\|S_\eta^{k+1}\right\|_{0,2,\mathcal{T}_\eta}^2\right).\end{aligned}$$

Inserting this estimate into (6.4.3), we infer that

$$\begin{aligned}E^{k+1} - E^k + \frac{\Delta t}{2} \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma \left|D\left(\sqrt{n^{k+1}}\right)_{K,\sigma}\right|^2 + \delta\,\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma \left|Dn_{K,\sigma}^{k+1}\right|^2 \\ \le C\left(\left\|n_\eta^{k+1}\right\|_{0,2,\mathcal{T}_\eta}^2 + \left\|S_\eta^{k+1}\right\|_{0,2,\mathcal{T}_\eta}^2\right).\end{aligned}$$

Summing this inequality over $k=0,\ldots,M_T-1$ and observing that the right-hand side is uniformly bounded, we conclude that $(\nabla^\eta n_\eta)$ is bounded in $L^2(\Omega_T)$, which finishes the proof. □

6.5 Convergence of the finite volume scheme

We prove Theorem 6.2.2. Consider the family $(n_\eta, S_\eta)_{\eta>0}$ of approximate solutions to (6.2.2)-(6.2.4). In order to apply compactness results, we need to control the difference $n_\eta(\cdot,t+\tau)-n_\eta(\cdot,t)$. To this end, let $\phi\in L^\infty(0,T;H^{2+\varepsilon}(\Omega))$, where $\varepsilon>0$. We denote by ϕ_K the average of ϕ in the control volume K. Using scheme (6.2.3) and the notation $n_\eta^{k+1} := n_\eta(\cdot,t^{k+1})$,

$$\begin{aligned}\sum_{K\in\mathcal{T}_\eta} \mathrm{m}(K)\,(n_K^{k+1}-n_K^k)\,\phi_K &\le \frac{\Delta t}{2}\sum_{K\in\mathcal{T}_\eta}\sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\left(\left|Dn_{K,\sigma}^{k+1}\right| + n_K^{k+1}\left|DS_{K,\sigma}^{k+1}\right|\right)|D\phi_{K,\sigma}| \\ &\le \Delta t\left(\left|n_\eta^{k+1}\right|_{1,2,\mathcal{T}_\eta}|\phi|_{1,2,\mathcal{T}_\eta} + \left\|n_\eta^{k+1}\right\|_{0,2,\mathcal{T}_\eta}\left|S_\eta^{k+1}\right|_{1,2,\mathcal{T}_\eta}|\phi|_{1,\infty,\mathcal{T}_\eta}\right) \\ &\le C\,\Delta t\left(\left|n_\eta^{k+1}\right|_{1,2,\mathcal{T}_\eta} + \left\|n_\eta^{k+1}\right\|_{0,2,\mathcal{T}_\eta}\left|S_\eta^{k+1}\right|_{1,2,\mathcal{T}_\eta}\right)\|\phi\|_{H^{2+\varepsilon}(\Omega)},\end{aligned}$$

where $C>0$ only depends on Ω. Summing over $k=0,\ldots,M_T-1$ and employing Hölder's inequality, the uniform bound on n_η from Proposition 6.4.2 and on S_η from (6.4.8) imply the existence of a constant $C>0$, independent of η, such that

$$\sum_{k=0}^{M_T-1}\sum_{K\in\mathcal{T}_\eta}\mathrm{m}(K)\left(n_K^{k+1}-n_K^k\right)\phi_K \le C\,\Delta t\,\|\phi\|_{L^\infty(0,T;H^{2+\varepsilon}(\Omega))}. \tag{6.5.1}$$

Now, similarly as in the proof of Lemma 10.6 in [85], for all $0<\tau<\Delta t$,

$$\begin{aligned}\int_0^{T-\tau}\int_\Omega &(n_\eta(x,t+\tau)-n_\eta(x,t))\,\phi(x,t)\,dx\,dt\\ &\le \sum_{k=0}^{M_T-1}\left(\int_0^{T-\tau}\chi_k(t,t+\tau)\,dt\right)\sum_{K\in\mathcal{T}_\eta}\mathrm{m}(K)\left(n_K^{k+1}-n_K^k\right)\phi_K,\end{aligned}$$

where

$$\chi_k(t,t+\tau)=\begin{cases}1 & \text{if } k\,\Delta t\in(t,t+\tau],\\ 0 & \text{if } k\,\Delta t\notin(t,t+\tau].\end{cases}$$

Inserting (6.5.1) into the above inequality and observing that

$$\int_0^{T-\tau}\chi_k(t,t+\tau)\,dt\le\tau\le\Delta t,$$

we infer that

$$\int_0^{T-\tau}\int_\Omega(n_\eta(x,t+\tau)-n_\eta(x,t))\,\phi(x,t)\,dx\,dt\le C\,\|\phi\|_{L^\infty(0,T;H^{2+\varepsilon}(\Omega))}.$$

This gives a uniform estimate for the time translations of n_η in $L^1(0,T;(H^{2+\varepsilon}(\Omega))')$. Since the embedding $H^1(\Omega)\hookrightarrow L^p(\Omega)$ is compact for all $1\le p<\infty$ in two space dimensions, we conclude from the discrete Aubin lemma [74] that there exists a subsequence of (n_η), not relabeled, such that, as $\eta\to 0$,

$$n_\eta\to n\quad\text{strongly in } L^2(0,T;L^p(\Omega)),\ p<\infty.$$

Furthermore, since $(\nabla^\eta n_\eta)$ is bounded in $L^2(\Omega_T)$, there exists $y\in L^2(\Omega_T)$ such that

$$\nabla^\eta n_\eta\rightharpoonup y\quad\text{weakly in } L^2(\Omega_T).$$

It is shown in the proof of Lemma 4.4 in [52] that $y=\nabla n$ in the sense of distributions. The bound of (S_η) in $L^2(0,T;H^1(\Omega))$ implies the existence of a subsequence, which is not relabeled, such that

$$S_\eta\rightharpoonup S,\quad \nabla^\eta S_\eta\rightharpoonup z\quad\text{weakly in } L^2(\Omega_T).$$

Again, it follows that $z=\nabla S$ in the sense of distributions.

The limit $\eta\to 0$ in the scheme (6.2.2)-(6.2.4) is performed exactly as in the proofs of Propositions 4.2 and 4.3 in [89], using the above convergence results and the fact that $(n_\eta\nabla^\eta S_\eta)$ converges weakly to $n\nabla S$ in $L^1(\Omega_T)$. Compared to [89], we have to pass to the limit also in the additional cross-diffusion term which does not give any difficulty since this term is linear in n_η. This shows that (n,S) solves the weak formulation (6.2.8)-(6.2.9), finishing the proof.

6.6 Long-time behavior

In this section, we prove Theorem 6.2.3. Similarly as in the proof of Proposition 6.4.1 (see (6.4.3) and (6.4.4)), we have

$$E[n^{k+1}|n^*] - E[n^k|n^*] = \sum_{K\in\mathcal{T}} \mathrm{m}(K)\,\left(H\left(n_K^{k+1}\right) - H\left(n_K^k\right)\right)$$

$$\leq -\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma \left|D\left(\sqrt{n^{k+1}}\right)_{K,\sigma}\right|^2 + \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma\, DS_{K,\sigma}^{k+1}\, Dn_{K,\sigma}^{k+1}. \tag{6.6.1}$$

In view of the identity $S^* = \mu\, n^*$, we can formulate the scheme (6.2.4) for all $K\in\mathcal{T}$ as

$$0 = \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\, D\left(S^{k+1} - S^*\right)_{K,\sigma} + \delta \sum_{\sigma\in\mathcal{E}_K} \tau_\sigma\, Dn_{K,\sigma}^{k+1} + \mathrm{m}(K)\,\left(\mu\,\left(n_K^{k+1} - n^*\right) - \left(S_K^{k+1} - S^*\right)\right).$$

Multiplying this equation by $(S_K^{k+1} - S^*)/\delta$ and summing over $K\in\mathcal{T}$ gives

$$\begin{aligned} 0 = &-\frac{1}{\delta} \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma \left|D\left(S^{k+1} - S^*\right)_{K,\sigma}\right|^2 - \sum_{\substack{\sigma\in\mathcal{E}_{\rm int},\\ \sigma=K|L}} \tau_\sigma\, DS_{K,\sigma}^{k+1}\, Dn_{K,\sigma}^{k+1} \\ &+ \frac{\mu}{\delta} \sum_{K\in\mathcal{T}} \mathrm{m}(K)\,\left(n_K^{k+1} - n^*\right)\,\left(S_K^{k+1} - S^*\right) - \frac{1}{\delta} \sum_{K\in\mathcal{T}} \mathrm{m}(K)\,\left(S_K^{k+1} - S^*\right)^2. \end{aligned}$$

Replacing the last term in (6.6.1) by the above equation and using the Cauchy-Schwarz and Young inequalities, it follows that

$$\begin{aligned} E[n^{k+1}|n^*] - E[n^k|n^*] + \Delta t\,\left|\sqrt{n^{k+1}}\right|_{1,2,\mathcal{T}}^2 + \frac{\Delta t}{\delta}\,\left\|S^{k+1} - S^*\right\|_{1,2,\mathcal{T}}^2 \\ = \frac{\mu\Delta t}{\delta} \sum_{K\in\mathcal{T}} \mathrm{m}(K)\,\left(n_K^{k+1} - n^*\right)\,\left(S_K^{k+1} - S^*\right) \\ \leq \frac{\mu^2\,\Delta t}{2\,\delta}\,\left\|n^{k+1} - n^*\right\|_{0,2,\mathcal{T}}^2 + \frac{\Delta t}{2\,\delta}\,\left\|S^{k+1} - S^*\right\|_{0,2,\mathcal{T}}^2. \end{aligned}$$

The second term on the right-hand side can be absorbed by the corresponding expression on the left-hand side. For the first term, we employ the continuous embedding of $BV(\Omega)$ into $L^2(\Omega)$ in dimension 2 (inequality (5.2.7) in Theorem 5.2.2) and the definition of the seminorm $|\cdot|_{1,1,\mathcal{T}}$:

$$\left\|n^{k+1} - n^*\right\|_{0,2,\mathcal{T}} \leq C(\Omega)\,\left|n^{k+1}\right|_{1,1,\mathcal{T}}.$$

Then, using inequality (6.4.9), we can estimate:

$$\left\|n^{k+1} - n^*\right\|_{0,2,\mathcal{T}}^2 \leq \frac{2}{\xi}\,C(\Omega)^2\,\|n_0\|_{L^1(\Omega)}\,\left|\sqrt{n^{k+1}}\right|_{1,2,\mathcal{T}}^2. \tag{6.6.2}$$

Setting $C^* = 2\,\mu^2\,C(\Omega)^2\,\|n_0\|_{L^1(\Omega)}/(\delta\,\xi)$, this yields

$$E[n^{k+1}|n^*] - E[n^k|n^*] + \Delta t\,(1 - C^*)\,\left|\sqrt{n^{k+1}}\right|_{1,2,\mathcal{T}}^2 + \frac{\Delta t}{2\,\delta}\,\left\|S^{k+1} - S^*\right\|_{1,2,\mathcal{T}}^2 \leq 0.$$

To proceed, we assume that $C^* < 1$. With the notation

$$I^{k+1} = (1 - C^*) \left|\sqrt{n^{k+1}}\right|_{1,2,\mathcal{T}}^2 + \frac{1}{2\,\delta} \left\|S^{k+1} - S^*\right\|_{1,2,\mathcal{T}}^2, \quad k \geq 0,$$

we have $E[n^{k+1}|n^*] + \Delta t\, I^{k+1} \leq E[n^k|n^*]$. Since H is convex, the entropy $E[n^k|n^*]$ is nonnegative. Therefore, for all $\ell \in \mathbb{N}$, summing over $k = 0, \ldots, \ell$,

$$0 \leq E[n^{\ell+1}|n^*] + \Delta t \sum_{k=0}^{\ell} I^{k+1} \leq E[n^0|n^*].$$

This means that the sequence $\sum_{k\in\mathbb{N}} I^{k+1}$ is finite and, since I^{k+1} is nonnegative, (I^{k+1}) converges to zero as $k \to \infty$. By definition of I^{k+1}, this shows that

$$\left|\sqrt{n^{k+1}}\right|_{1,2,\mathcal{T}} \to 0, \quad \left\|S^{k+1} - S^*\right\|_{1,2,\mathcal{T}} \to 0 \quad \text{as } k \to \infty$$

and, taking into account (6.6.2),

$$\left\|n^{k+1} - n^*\right\|_{0,2,\mathcal{T}} \to 0 \quad \text{as } k \to \infty,$$

which concludes the proof.

6.7 Numerical experiments

In this section, we investigate the numerical convergence rates and give some examples illuminating the long-time behavior of the finite volume solutions to nonhomogeneous steady states.

6.7.1 Numerical convergence rates

We compute first the spatial convergence rate of the numerical scheme. We consider the system (6.1.1) on the square $\Omega = (-\frac{1}{2}, \frac{1}{2})^2$. The time step is chosen to be $\Delta t = 10^{-8}$, and the final time is $t = 10^{-4}$. The initial data is the Gaussian

$$n_0(x, y) = \frac{M}{2\pi\theta} \exp\left(-\frac{(x - x_0)^2 + (y - y_0)^2}{2\theta}\right),$$

where $\theta = 10^{-2}$, $M = \|n_0\|_{L^1(\Omega)} = 6\pi$, and $x_0 = y_0 = 0.1$. The model parameters are $\delta = 10^{-3}$ and $\mu = 1$. We compute the numerical solution on a sequence of square meshes. The coarsest mesh is composed of 4×4 squares. The sequence of meshes is obtained by dividing successively the size of the squares by 4. Then, the finest grid is made of 256×256 squares. The relative L^p error at time t is given by

$$e_{\Delta x} = \frac{\|n_{\Delta x}(\cdot, t) - n_{\mathrm{ex}}(\cdot, t)\|_{L^p(\Omega)}}{\|n_{\mathrm{ex}}(\cdot, t)\|_{L^p(\Omega)}},$$

where $n_{\Delta x}$ represents the approximation of the cell density computed from a mesh of size Δx and n_{ex} is the "exact" solution computed from a mesh with 256×256 squares and with time step size $\Delta t = 10^{-8}$. The numerical scheme is said to be of order m if for all sufficiently small $\Delta x > 0$, it holds that $e_{\Delta x} \leq C(\Delta x)^m$ for some constant $C > 0$. Figure 6.1 shows that the convergence rates in the L^1, L^2, and L^∞ norms are around one. As expected, the scheme is of first order.

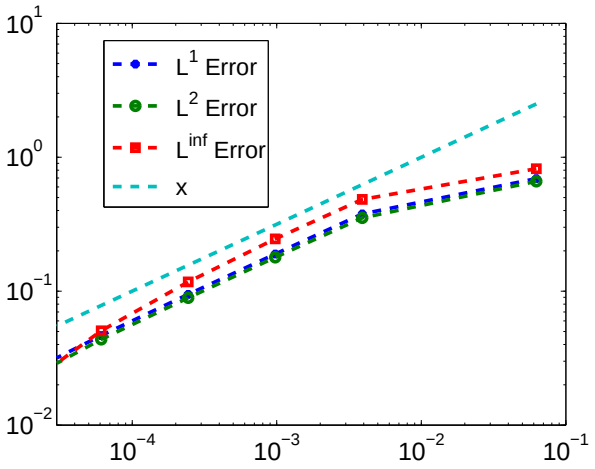


Figure 6.1: Spatial convergence orders in the L^1, L^2 and L^∞ norms.

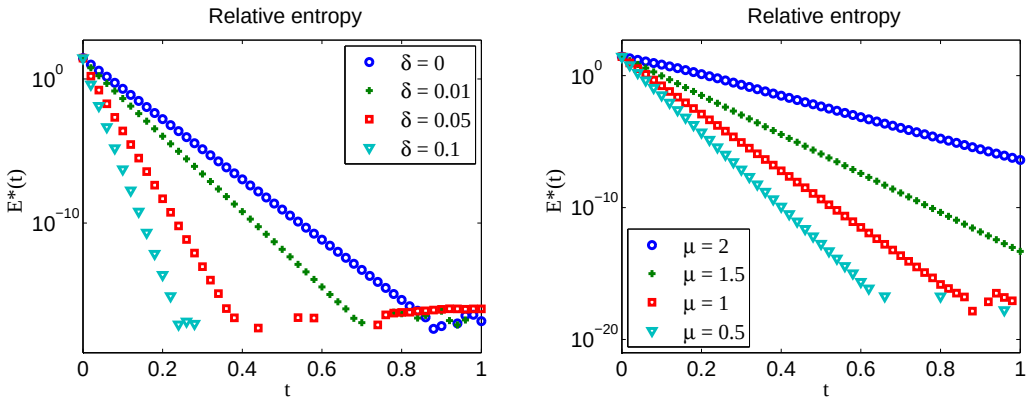
6.7.2 Decay rates

According to Theorem 6.2.3, the solution to the Keller-Segel system converges to the homogeneous steady state if μ or $1/\delta$ are sufficiently small. We will verify this property experimentally. To this end, let $\Omega = (-\frac{1}{2}, \frac{1}{2})^2$ and

$$n_0(x, y) = \frac{M}{2\pi\theta} \exp\left(-\frac{x^2 + y^2}{2\theta}\right),$$

where $\theta = 10^{-2}$ and $M = 5\pi$. We compute the approximate solution on a 32×32 Cartesian grid, and we choose $\Delta t = 2 \cdot 10^{-4}$. In Figure 6.2, we depict the temporal evolution of the relative entropy $E^*(t^k) = E[n^k | n^*]$ in semi-logarithmic scale. In all cases shown, the convergence seems to be of exponential rate. The rate becomes larger for larger values of δ or smaller values of μ. This is in agreement with the theoretical result: indeed, it is shown in [111, Theorem 1.3] that the exponential decay rate is proportional to $1 - C(\Omega)\,\mu^2 \|n_0\|_{L^1(\Omega)}/\delta$, and hence, this rate becomes larger for smaller μ or $1/\delta$.

As a numerical check, we computed the evolution of the relative entropies for different grid sizes N and different time step sizes Δt. Figure 6.3 shows that the decay rate does not depend on the time step or the mesh considered.



(a) $\mu = 1$. (b) $\delta = 10^{-3}$.

Figure 6.2: Relative entropy $E[n^k|n^*]$ versus time t^k in semi-logarithmic scale for various values of δ and μ.

6.7.3 Nonsymmetric initial data on a square

In this subsection, we explore the behavior of the solutions to (6.1.1) for different values of δ. We choose $\Omega = (-\frac{1}{2}, \frac{1}{2})^2$ with a 64×64 Cartesian grid, $\mu = 1$, and $\Delta t = 2 \cdot 10^{-5}$. We consider two nonsymmetric initial functions with mass 6π:

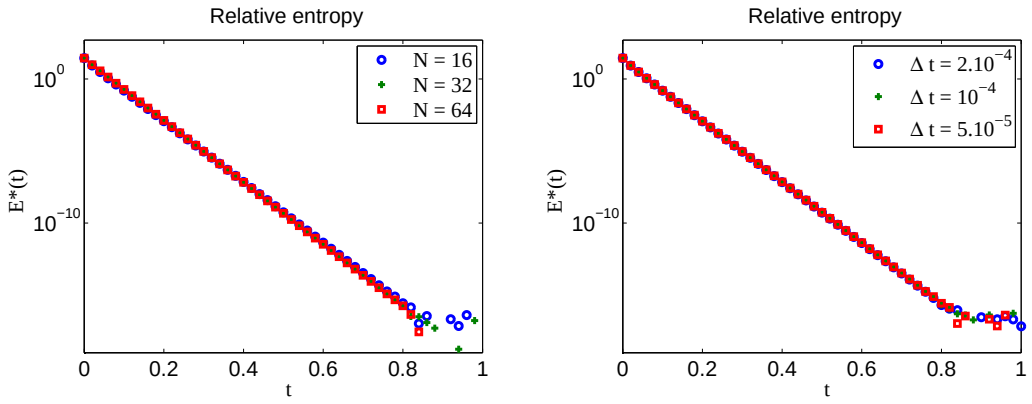
$$n_{0,1}(x,y) = \frac{6\pi}{2\pi\theta} \exp\left(-\frac{(x-x_0)^2 + (y-y_0)^2}{2\theta}\right), \tag{6.7.1}$$

$$n_{0,2}(x,y) = \frac{4\pi}{2\pi\theta} \exp\left(-\frac{(x-x_0)^2 + (y-y_0)^2}{2\theta}\right) + \frac{2\pi}{2\pi\theta} \exp\left(-\frac{(x-x_1)^2 + (y-y_1)^2}{2\theta}\right), \tag{6.7.2}$$

where $\theta = 10^{-2}$, $x_0 = y_0 = 0.1$, and $x_1 = y_1 = -0.2$ (see Figure 6.4).

We consider first the case $\delta = 0$, which corresponds to the classical parabolic-elliptic Keller-Segel system. In this case, our finite volume scheme coincides with that of [89]. We recall that solutions to the classical parabolic-elliptic model blow up in finite time if the initial mass satisfies $M > 4\pi$ [142] (in the non-radial case). The numerical results at a time just before the numerical blow-up are presented in Figure 6.5. We observe the blow-up of the cell density in finite time, and the blow-up occurs at the boundary, as expected. More precisely, it occurs at that corner which is closest to the global maximum of the initial datum.

Next, we choose $\delta = 10^{-3}$ and $\delta = 10^{-2}$. According to Theorem 6.2.1, the numerical solution exists for all time. This behavior is confirmed in Figure 6.6, where we show the cell density at time $t = 5$. At this time, the solution is very close to the steady state which is nonhomogeneous. We observe a smoothing effect of the cross-diffusion parameter δ; the cell density maximum decreases with increasing values of δ.



(a) $\Delta t = 2.10^{-5}$. (b) $N = 16$.

Figure 6.3: Relative entropy $E[n^k|n^*]$ versus time t^k in semi-logarithmic scale for various mesh and time step sizes.

6.7.4 Symmetric initial data on a square

We consider, as in the previous subsection, the domain $\Omega = (-\frac{1}{2}, \frac{1}{2})^2$ with a 64×64 Cartesian grid, $\mu = 1$, and $\Delta t = 10^{-5}$. Here, we consider the radially symmetric initial datum

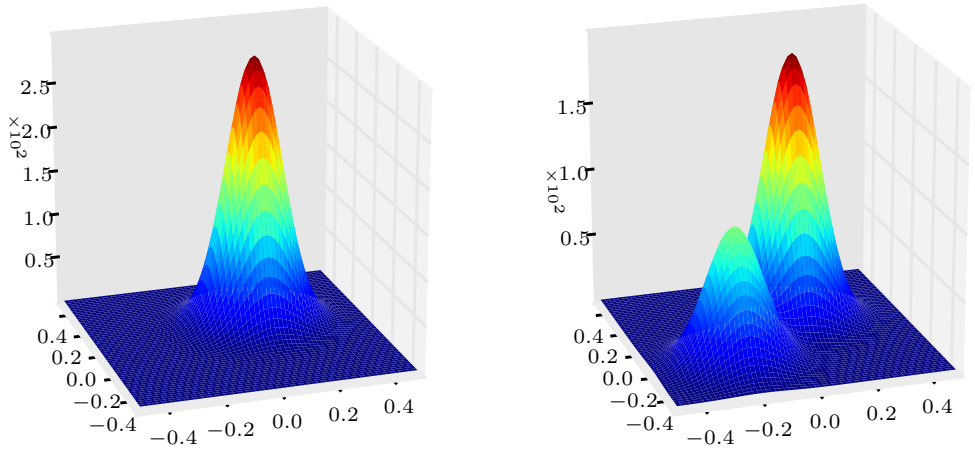
$$n_{0,3}(x, y) = \frac{M}{2\pi\theta} \exp\left(-\frac{x^2 + y^2}{2\theta}\right) \tag{6.7.3}$$

with $M = 20\pi$ and $\theta = 10^{-2}$. Since $M > 8\pi$ and the initial datum is radially symmetric, we expect that the solution to the classical Keller-Segel model ($\delta = 0$) blows up in finite time [141, 150]. Figure 6.7 shows that this is indeed the case, and blow-up occurs in the center of the domain.

In contrast to the classical Keller-Segel system, when taking $\delta = 10^{-3}$, the cell density peak moves to a corner of the domain and converges to a nonhomogeneous steady state (see Figure 6.8). The time evolution of the L^∞ norm of the cell density shows an interesting behavior (see Figure 6.9). We observe two distinct levels. The first one is reached almost instantaneously. The L^∞ norm stays almost constant and the cell density seems to stabilize at an intermediate symmetric state (Figure 6.8(a)). After some time, the L^∞ norm increases sharply and the cell density peak moves to the boundary (Figure 6.8(b)). Then the solution stabilizes again (Figure 6.8(c)). We note that we obtain the same steady state when using a Gaussian centered at $(-10^{-3}, -10^{-3})$.

6.7.5 Nonsymmetric initial data on a rectangle

We consider the domain $\Omega = (-1, 1) \times (-\frac{1}{2}, \frac{1}{2})$ and compute the approximate solutions on a 128×64 Cartesian grid with $\Delta t = 5 \cdot 10^{-5}$. The secretion rate is again $\mu = 1$, and we choose the initial data $n_{0,1}$ and $n_{0,2}$, defined in (6.7.1)-(6.7.2) with mass $M = 6\pi$. If



(a) Initial datum $n_{0,1}$.

(b) Initial datum $n_{0,2}$.

Figure 6.4: Initial cell densities.

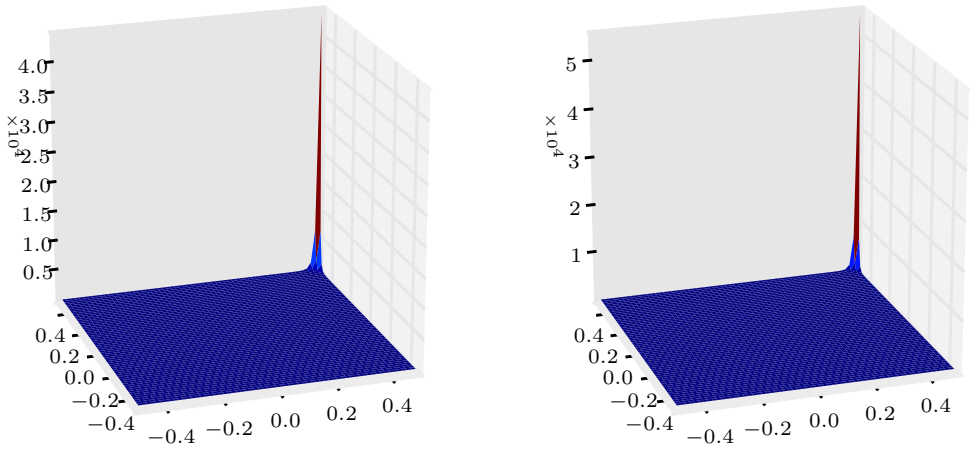
$\delta = 0$, the solution blows up in finite time and the blow up occurs in a corner as in the square domain (see Figure 6.10). If $\delta = 10^{-3}$, the approximate solutions converge to a non-homogeneous steady state (Figure 6.11). Interestingly, before moving to the corner, the solution corresponding to the nonsymmetric initial datum $n_{0,2}$ shows some intermediate behavior; see Figure 6.11(b).

6.7.6 Symmetric initial data on a rectangle

The domain is still the rectangle $\Omega = (-1, 1) \times (-\frac{1}{2}, \frac{1}{2})$, we take a 128×64 Cartesian grid, $\mu = 1$, and $\Delta t = 10^{-5}$. We choose the initial datum $n_{0,3}$, defined in (6.7.3), with $M = 20\pi$. Clearly, the approximate solution to the classical Keller-Segel model $\delta = 0$ blows up in finite time in the center $(0, 0)$ of the rectangle. When $\delta = 10^{-3}$, the behavior is different and similar to that in a square domain. However, in contrast to the square domain, the cell density peak first moves straight to the closest boundary point before moving to a corner of the domain (Figure 6.12). This behavior is reflected in the time evolution of the L^∞ norm; see Figure 6.13: there exist two intermediate states, one up to time $t \approx 0.9$ and another in the interval $(0.9, 2.3)$, and one final state for long times. We note that the same qualitative behavior is obtained using $\delta = 10^{-2}$.

6.8 Conclusion

In this chapter, we analyse a finite volume scheme for a parabolic-elliptic Keller-Segel model with cross-diffusion. We prove the convergence of the discrete solution to the continuous one when the discretization parameters tend to zero, and we investigate the

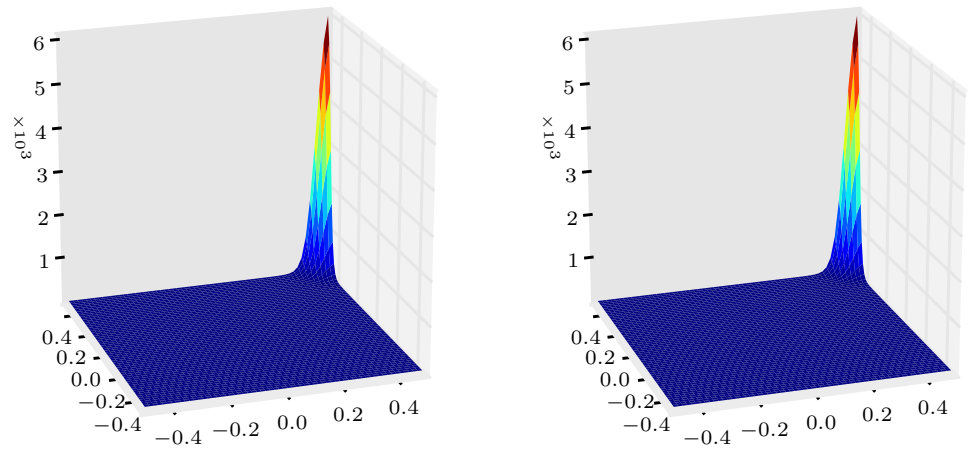


(a) Cell density computed from $n_{0,1}$, $t = 1$. (b) Cell density computed from $n_{0,2}$, $t = 0.6$.

Figure 6.5: Cell density computed from nonsymmetric initial data with $M = 6\pi$ and $\delta = 0$.

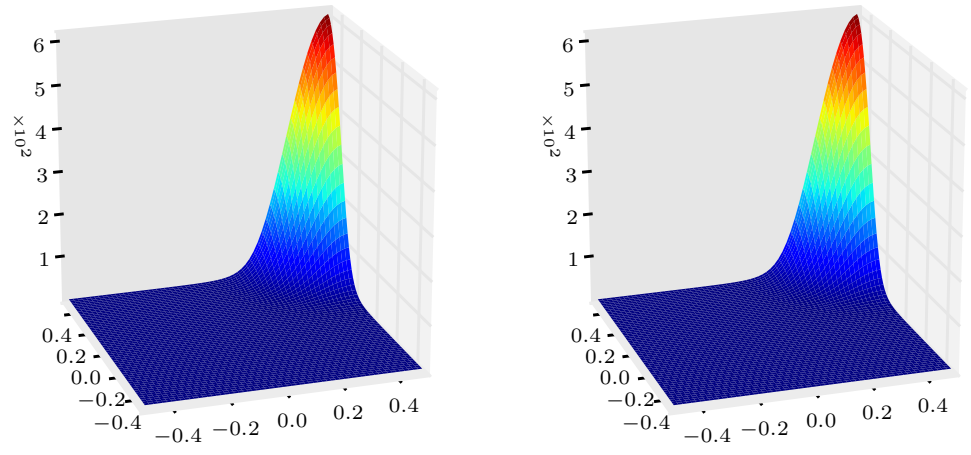
long-time behavior of the approximate solution. The proof of both results is based on a discrete version of an entropy–dissipation relation.

We also present some numerical results illuminating the long-time behavior of the solutions. On the one hand, for sufficiently large values of the diffusion parameter δ, we observe an exponential decay to the homogeneous steady-state. A future work would be to prove this exponential decay rate, and to this end we need to establish a discrete logarithmic Sobolev inequality. On the other hand, the numerical solutions seem to converge to nonhomogeneous steady-states for intermediate values of δ.



(a) Cell density computed from $n_{0,1}$, $\delta = 10^{-3}$.

(b) Cell density computed from $n_{0,2}$, $\delta = 10^{-3}$.



(c) Cell density computed from $n_{0,1}$, $\delta = 10^{-2}$.

(d) Cell density computed from $n_{0,2}$, $\delta = 10^{-2}$.

Figure 6.6: Cell density computed at $t = 5$ from nonsymmetric initial data with $M = 6\pi$ for different values of δ.

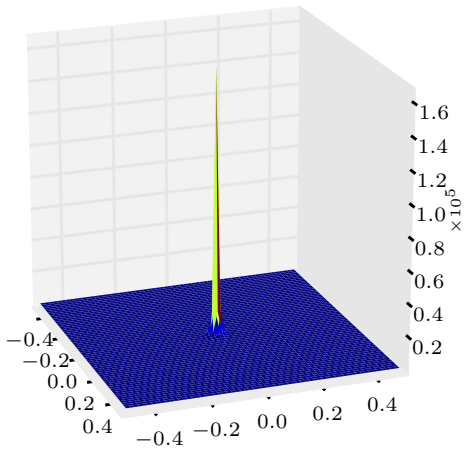
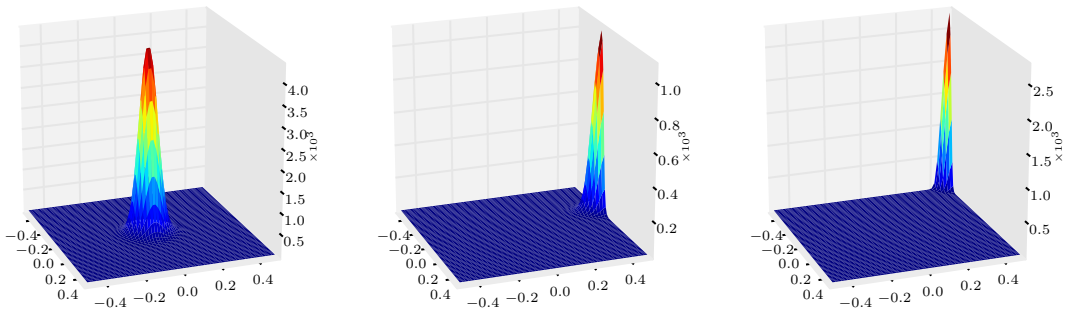


Figure 6.7: Cell density at time $t = 0.05$ computed from the radially symmetric initial datum $n_{0,3}$ with $M = 20\pi$ and $\delta = 0$.



(a) $t = 0.6$. (b) $t = 0.73$. (c) $t = 5$.

Figure 6.8: Cell density computed from the radially symmetric initial datum $n_{0,3}$ with $M = 20\pi$ and $\delta = 10^{-3}$.

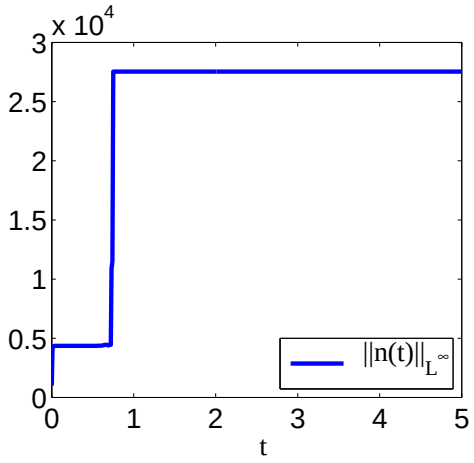
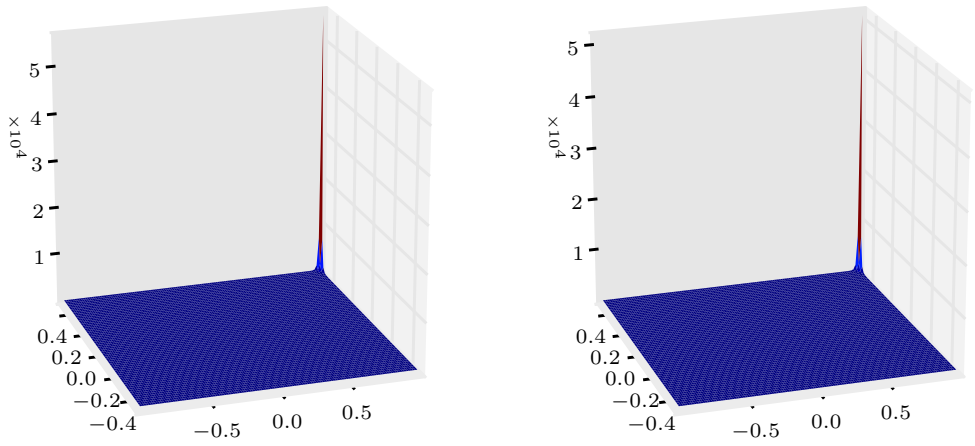
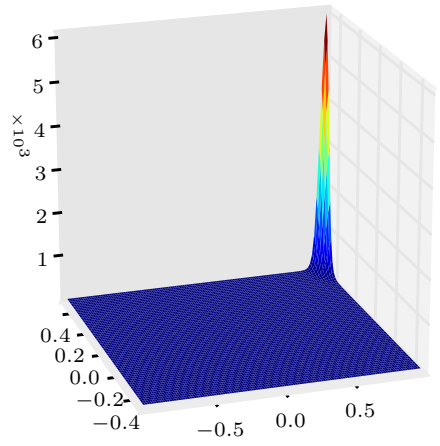


Figure 6.9: Time evolution of $\|n^k\|_{L^\infty(\Omega)}$ computed from the radially symmetric initial datum $n_{0,3}$ with $M = 20\pi$ and $\delta = 10^{-3}$.

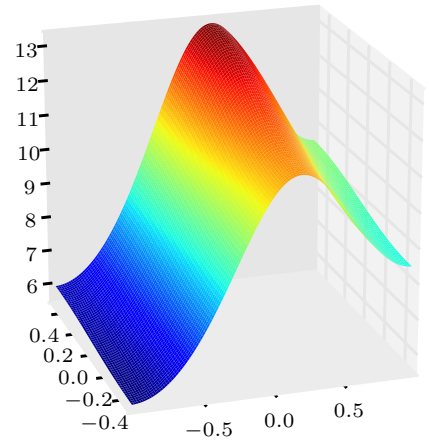


(a) Cell density computed from $n_{0,1}$, $t = 0.5$. (b) Cell density computed from $n_{0,2}$, $t = 1.7$.

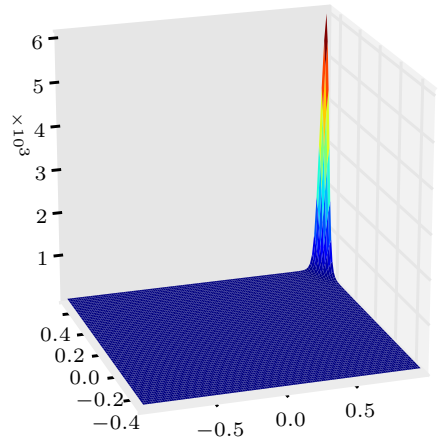
Figure 6.10: Cell density computed from nonsymmetric initial data with $M = 6\pi$ and $\delta = 0$.



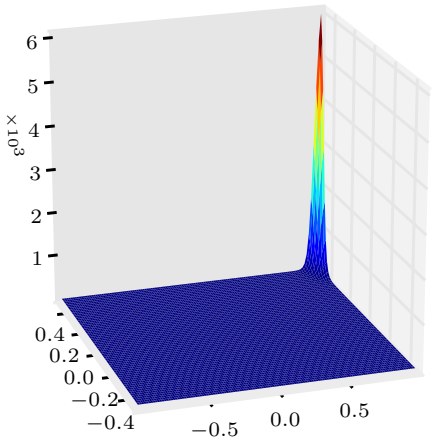
(a) Cell density computed from $n_{0,1}$, $t = 1$.



(b) Cell density computed from $n_{0,2}$, $t = 1$.

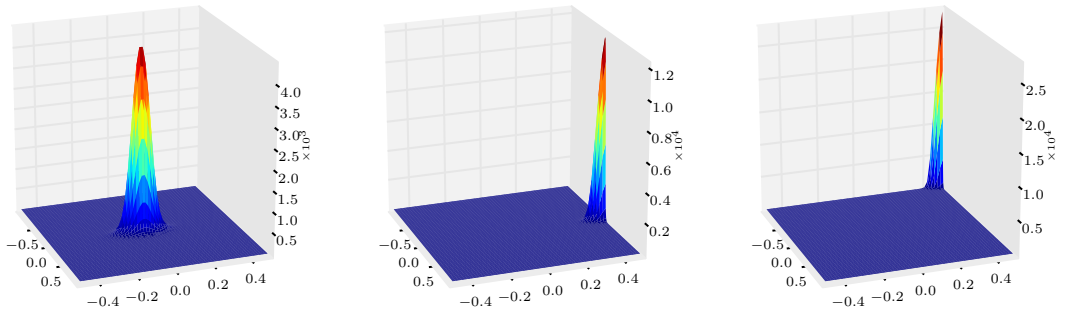


(c) Cell density computed from $n_{0,1}$, $t = 5$.



(d) Cell density computed from $n_{0,2}$, $t = 5$.

Figure 6.11: Cell density computed from nonsymmetric initial data with $M = 6\pi$ and $\delta = 10^{-3}$.



(a) $t = 0.7$. (b) $t = 1.9$. (c) $t = 2.5$.

Figure 6.12: Cell density computed from the symmetric initial datum $n_{0,3}$ with $M = 20\pi$ and $\delta = 10^{-3}$.

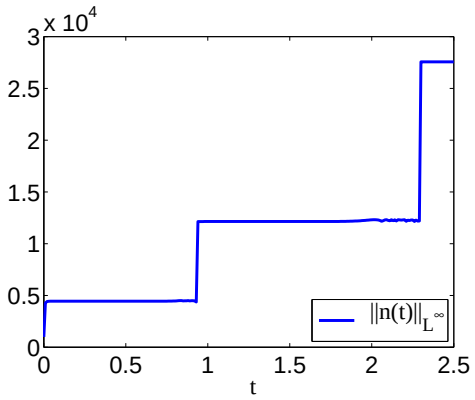


Figure 6.13: Time evolution of $\|n^k\|_{L^\infty(\Omega)}$ computed from the radially symmetric initial datum $n_{0,3}$ with $M = 20\pi$ and $\delta = 10^{-3}$.

Conclusion et perspectives

Dans ce dernier chapitre, nous montrons comment les travaux effectués dans cette thèse peuvent être approfondis en dégageant quelques pistes de recherches possibles.

1 Partie I

Dans cette partie, nous avons étudié un schéma complètement implicite en temps, de type volumes finis en espace et défini avec le flux de Scharfetter-Gummel. Ce schéma est appliqué au système de dérive-diffusion linéaire pour les semi-conducteurs. D'une part, nous avons obtenu une preuve complète de la convergence en temps long des solutions numériques vers une approximation de l'équilibre thermique. D'autre part, nous avons considéré un autre type d'asymptotique, la limite quasi-neutre, et nous avons démontré la convergence du schéma pour toute valeur du paramètre $\lambda > 0$. Dans ces deux études, le point clé est une estimation d'énergie–dissipation discrète, obtenue grâce aux propriétés particulières du flux de Scharfetter-Gummel et au caractère implicite en temps du schéma. Plusieurs questions restent cependant ouvertes concernant l'étude de ces deux asymptotiques.

1.1 Hypothèses moins restrictives sur le dopage

Les résultats obtenus dans la partie I le sont sous l'hypothèse restrictive d'un profil de dopage nul, peu réaliste d'un point de vue physique. De plus, nous présentons un certain nombre de résultats numériques qui semblent indiquer que cette hypothèse n'est pas nécessaire. Au niveau continu, l'hypothèse d'un dopage nul est également formulée par I. Gasser [94] et A. Jüngel et Y. J. Peng [121] dans leur étude du passage à la limite quasi-neutre. Par contre, I. Gasser, C. D. Levermore, P. Markowich et C. Schmeiser [95] étudient cette limite en supposant uniquement que $C \in L^\infty(\Omega) \cap H^1(\Omega)$, et que le signe de C est constant sur Ω. Cependant, ils considèrent seulement le cas de conditions au bord de Neumann homogènes, ce qui simplifie l'étude. Concernant l'étude de l'asymptotique en temps long au niveau continu, H. Gajewski et K. Gärtner [92] supposent que $C \in L^\infty(\Omega)$, et c'est aussi l'hypothèse formulée par A. Jüngel [119] pour l'étude du système de dérive-diffusion non linéaire. Il faudrait donc essayer d'étendre les résultats du chapitre 1 au cas d'un dopage $C \in L^\infty$ et tenter également d'obtenir ceux du chapitre 2 en faisant par exemple dans un premier temps l'hypothèse d'un dopage constant.

1.2 Démonstration complète du caractère AP du schéma

Dans le chapitre 2, nous avons démontré la convergence du schéma de Scharfetter-Gummel implicite pour toute valeur de λ strictement positive. Cette étude constitue la première étape pour montrer que le schéma préserve l'asymptotique pour la limite quasi-neutre. Il faut à présent étudier la limite $\lambda \to 0$ lorsque le paramètre de discrétisation δ est fixé dans le schéma (ce qui correspond à la flèche $\mathcal{P}_\delta^\lambda \longrightarrow \mathcal{P}_\delta^0$ quand $\lambda \to 0$ dans le diagramme de la page 7) et vérifier que le schéma limite obtenu $\mathcal{P}_\delta^0$ fournit bien une approximation du problème limite continu ($\mathcal{P}^0$) (ce qui correspond à la flèche $\mathcal{P}_\delta^0 \longrightarrow \mathcal{P}^0$ quand $\delta \to 0$). Ce travail, en collaboration avec C. Chainais-Hillairet et M.-H. Vignal, est actuellement en cours.

2 Partie II

Dans la partie II, nous avons proposé deux nouveaux schémas volumes finis pour des problèmes de convection-diffusion non linéaires, avec pour objectif la préservation du comportement en temps long au niveau discret. Dans le chapitre 3, nous avons généralisé le schéma de Scharfetter-Gummel au cas d'une diffusion non linéaire, et nous en avons étudié la convergence dans le cas non dégénéré. Nous avons appliqué ce nouveau schéma à l'équation des milieux poreux et au système de dérive-diffusion non linéaire, et obtenu des résultats satisfaisants concernant l'asymptotique en temps long. Dans le chapitre 4, nous avons proposé une nouvelle discrétisation spatiale, valable pour des problèmes paraboliques non linéaires assez généraux. Ce nouveau schéma préserve non seulement l'asymptotique en temps long, mais reste d'ordre deux en espace même dans le cas dégénéré. Plusieurs points pourraient encore être approfondis concernant l'étude de ces deux schémas.

Les résultats du chapitre 3 sont démontrés sous des hypothèses assez fortes. Il faudrait d'abord regarder si la convergence du schéma peut être prouvée dans le cas où l'équation dégénère en certains points. Dans cet objectif, la première étape serait de démontrer l'estimation $L^2(0,T;H^1)$ (Proposition 3.3.2) sans utiliser la borne inférieure uniforme sur u_δ. Toujours concernant le chapitre 3, l'hypothèse $\mathrm{div}(\mathbf{q}) = 0$ est, elle aussi, très forte. Elle nous permet de faire une étude générale du schéma, mais il serait intéressant de regarder plus précisément ce qui arrive quand on applique le schéma à l'équation des milieux poreux (2.10), qui correspond au cas $\mathbf{q} = x$, et au système de dérive-diffusion (2.11), qui correspond à $\mathbf{q} = \nabla V$. Par rapport à ce système, il semble que les résultats peuvent être obtenus en faisant l'hypothèse d'un dopage nul sur Ω. Enfin, nous pourrions établir une estimation d'entropie–dissipation discrète pour ce schéma, en utilisant des techniques similaires à celles présentées dans le chapitre 1. Toutefois, il semble pour cela nécessaire de considérer une discrétisation entièrement implicite en temps.

Nous avons prouvé dans le chapitre 4 une estimation d'entropie–dissipation uniquement pour le cas d'une équation avec convection linéaire (2.6). Nous pourrions nous intéresser au cas d'une convection non linéaire, en considérant l'exemple de l'équation (2.15) pour les bosons et les fermions. Pouvons-nous alors obtenir une analogue discret de l'inégalité

d'entropie–dissipation (2.5) ? Il faut également étudier plus précisément le comportement du schéma dans les cas où la solution explose en temps fini (par exemple le cas tridimensionnel pour les bosons).

Enfin, dans une perspective plus générale, nous pourrions tenter d'obtenir des résultats concernant la vitesse de convergence vers l'équilibre des solutions approchées. Nous observons numériquement pour l'équation (2.6) un taux de convergence exponentiel conforme au résultat démontré dans [45]. La preuve de ce fait au niveau continu est fondée sur une dérivation par rapport au temps de la dissipation d'énergie, dont il faudrait essayer d'obtenir un analogue au niveau discret.

3 Partie III

Dans la partie III enfin, nous avons analysé un schéma volumes finis pour un modèle de Keller-Segel avec diffusion croisée, dans le cas parabolique-elliptique. D'une part, nous avons prouvé la convergence de la solution approchée vers la solution faible du problème continu quand les paramètres de discrétisation tendent vers zéro. D'autre part, nous avons étudié la convergence en temps long de la solution numérique vers l'état stationnaire homogène, dans le cas particulier où le paramètre de diffusion δ est suffisamment grand ou le taux de sécrétion μ est suffisamment petit. Ces deux études sont fondées sur l'obtention d'estimations a priori obtenues grâce à l'utilisation de versions discrètes d'inégalités fonctionnelles, que nous avons établies dans un contexte assez général.

Étude du schéma pour le système parabolique-parabolique

Au niveau continu, S. Hittmeir et A. Jüngel étudient le système de Keller-Segel avec diffusion croisée (3.10)–(3.12) aussi bien dans le cas parabolique-elliptique que dans le cas parabolique-parabolique. Nous pouvons considérer une discrétisation volumes finis implicite pour le modèle parabolique-parabolique analogue à celle présentée dans le chapitre 6 pour le modèle parabolique-elliptique. En utilisant les mêmes techniques, nous obtenons alors l'existence de solutions numériques, ainsi qu'une estimation d'entropie discrète. Cependant, dans ce cas-là nous n'obtenons qu'une borne uniforme dans $L^2(0,T;W^{1,1}(\Omega))$ pour n et non dans $L^2(0,T;H^1(\Omega))$. Au niveau continu, cette borne uniforme est suffisante pour mener l'étude du système parabolique-parabolique mais dans notre version discrétisée, un terme en $\nabla n \cdot \nabla S$ apparaît pour l'instant dans le passage à la limite et ne peut plus être traité comme dans le cas parabolique-elliptique du fait de ce défaut d'estimation a priori. Il faudrait essayer de trouver une autre manière de passer à la limite dans le schéma pour éviter cet écueil et obtenir ainsi l'étude complète du schéma dans le cas parabolique-parabolique.

Taux de convergence vers l'état stationnaire homogène

Concernant la convergence en temps long des solutions vers l'état stationnaire homogène dans le cas où μ est suffisamment petit ou δ est suffisamment grand, S. Hittmeir

et A. Jüngel obtiennent au niveau continu un résultat plus précis, à savoir que la vitesse de convergence est exponentielle [111, Theorem 1.3]. Nos résultats numériques sont en accord avec ce résultat, cependant, nous ne démontrons pas pour l'instant ce taux de convergence exponentiel au niveau discret. En effet, la preuve au niveau continu nécessite l'utilisation de l'inégalité de Sobolev logarithmique suivante :

$$\int_\Omega n \log(n)\, dx \leq C \int_\Omega \left|\nabla \sqrt{n}\right|^2 dx, \tag{3.1}$$

qui permet d'obtenir un contrôle de la dérivée temporelle de l'entropie relative $E[n(t)|n^*]$ (3.14) par $E[n(t)|n^*]$ et d'en déduire la décroissance exponentielle de cette quantité par application du lemme de Gronwall. Nous ne disposons pas pour l'instant de version discrète de l'inégalité de Sobolev logarithmique (3.1). Ce travail pourrait s'inscrire dans la continuité de celui entamé au chapitre 5 où des analogues discrets d'inégalités fonctionnelles classiques dans le contexte continu sont démontrés.

Bibliographie

[1] R. A. Adams. *Sobolev spaces*. Academic Press, New York-London, 1975. Pure and Applied Mathematics, Vol. 65. 26, 151, 156, 164

[2] H. W. Alt, S. Luckhaus, and A. Visintin. On nonstationary flow through porous media. *Ann. Mat. Pura Appl. (4)*, 136(1) :303–316, 1984. 91

[3] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems*. Oxford Mathematical Monographs. The Clarendon Press Oxford University Press, New York, 2000. 29, 155

[4] B. Andreianov, M. Bendahmane, and R. Ruiz Baier. Analysis of a finite volume method for a cross-diffusion model in population dynamics. *Math. Model. Meth. Appl. Sci.*, 21(2) :307–344, 2011. 27, 151

[5] B. Andreianov, F. Boyer, and F. Hubert. Discrete duality finite volume schemes for Leray-Lions-type elliptic problems on general 2D meshes. *Numer. Methods Partial Differential Equations*, 23(1) :145–195, 2007. 28, 148, 166, 168

[6] P. Angot, V. Dolejší, M. Feistauer, and J. Felcman. Analysis of a combined barycentric finite volume-nonconforming finite element method for nonlinear convection-diffusion problems. *Appl. Math.*, 43 :263–310, 1998. 18

[7] T. Arbogast, M. F. Wheeler, and N. Y. Zhang. A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media. *SIAM J. Numer. Anal.*, 33(4) :1669–1687, 1996. 18, 119

[8] D. Aregba-Driollet, R. Natalini, and S. Tang. Explicit diffusive kinetic schemes for nonlinear degenerate parabolic systems. *Math. Comp.*, 73(245) :63–94 (electronic), 2004. 18, 119

[9] F. Arimburgo, C. Baiocchi, and L. D. Marini. Numerical approximation of the 1-D nonlinear drift-diffusion model in semiconductors. In *Nonlinear kinetic theory and mathematical aspects of hyperbolic systems (Rapallo, 1992)*, volume 9 of *Ser. Adv. Math. Appl. Sci.*, pages 1–10. World Sci. Publ., River Edge, NJ, 1992. 7, 89

[10] A. Arnold, J. A. Carrillo, L. Desvillettes, J. Dolbeault, A. Jüngel, C. Lederman, P. A. Markowich, G. Toscani, and C. Villani. Entropies and equilibria of many-particle systems : an essay on recent research. *Monatsh. Math.*, 142(1-2) :35–43, 2004. 3, 4

[11] A. Arnold and A. Unterreiter. Entropy decay of discretized Fokker-Planck equations. I. Temporal semidiscretization. *Comput. Math. Appl.*, 46(10-11) :1683–1690, 2003. 18, 19, 123

[12] D. N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19 :742–760, 1982. 28, 152

[13] J. W. Barrett and P. Knabner. Finite element approximation of the transport of reactive solutes in porous media. II. Error estimates for equilibrium adsorption processes. *SIAM J. Numer. Anal.*, 34(2) :455–479, 1997. 18, 119

[14] C. Bataillon, F. Bouchon, C. Chainais-Hillairet, J. Fuhrmann, E. Hoarau, and R. Touzani. Numerical methods for the simulation of a corrosion model with moving oxide layer. Preprint 2012. 60

[15] N. Ben Abdallah, I. M. Gamba, and G. Toscani. On the minimization problem of sub-linear convex functionals. *Kinet. Relat. Models*, 4(4) :857–871, 2011. 17, 18, 123, 140

[16] M. Bessemoulin-Chatard. A finite volume scheme for convection-diffusion equations with nonlinear diffusion derived from the Scharfetter-Gummel scheme. To appear in Numerische Mathematik. xiii, 85, 119, 123

[17] M. Bessemoulin-Chatard, C. Chainais-Hillairet, and F. Filbet. On discrete functional inequalities for some finite volume schemes. Preprint 2012. xiv, 149

[18] M. Bessemoulin-Chatard, C. Chainais-Hillairet, and M.-H. Vignal. Convergence of a fully implicit scheme for the drift-diffusion system. Stability at the quasineutral limit. Preprint 2012. xiii, 55

[19] M. Bessemoulin-Chatard and F. Filbet. A finite volume scheme for nonlinear degenerate parabolic equations. To appear in SIAM Journal on Scientific Computing. xiii, 117

[20] M. Bessemoulin-Chatard and A. Jüngel. A finite volume scheme for a Keller-Segel model with additional cross-diffusion. Preprint 2012. xiv, 175

[21] A. Blanchet, V. Calvez, and J. A. Carrillo. Convergence of the mass-transport steepest descent scheme for the subcritical Patlak-Keller-Segel model. *SIAM J. Numer. Anal.*, 46(2) :691–721, 2008. 31, 178

[22] A. Blanchet, E. A. Carlen, and J. A. Carrillo. Functional inequalities, thick tails and asymptotics for the critical mass Patlak-Keller-Segel model. *J. Funct. Anal.*, 262(5) :2142–2230, 2012. 24, 176

[23] A. Blanchet, J. A. Carrillo, and N. Masmoudi. Infinite time aggregation for the critical Patlak-Keller-Segel model in $\mathbb{R}^2$. *Communications on Pure and Applied Mathematics*, 61(10) :1449–1481, 2008. 24, 177

[24] A. Blanchet, J. Dolbeault, and B. Perthame. Two-dimensional Keller-Segel model : optimal critical mass and qualitative properties of the solutions. *Electron. J. Differential Equations*, 2006 :No. 44, 32 pp. (electronic), 2006. 24, 177

[25] F. Bouchut, R. Eymard, and A. Prignet. Finite volume schemes for the approximation via characteristics of linear convection equations with irregular data. *J. Evol. Equ.*, 11(3) :687–724, 2011. 27, 28, 151, 152

[26] F. Bouchut, F. R. Guarguaglini, and R. Natalini. Diffusive BGK approximations for nonlinear multidimensional parabolic equations. *Indiana Univ. Math. J.*, 49 :723–749, 2000. 18

[27] F. Boyer and F. Hubert. Finite volume method for 2D linear and nonlinear elliptic problems with discontinuities. *SIAM J. Numer. Anal.*, 46(6) :3032–3070, 2008. 29, 166

[28] Y. Brenier and E. Grenier. Limite singulière du système de Vlasov-Poisson dans le régime de quasi neutralité : le cas indépendant du temps. *C. R. Acad. Sci. Paris Sér. I Math.*, 318(2) :121–124, 1994. 6

[29] M. P. Brenner, P. Constantin, L. P. Kadanoff, A. Schenkel, and S. C. Venkataramani. Diffusion, attraction and collapse. *Nonlinearity*, 12(4) :1071–1098, 1999. 24

[30] S. C. Brenner. Poincaré-Friedrichs inequalities for piecewise H^1 functions. *SIAM J. Numer. Anal.*, 41(1) :306–324 (electronic), 2003. 28, 152

[31] H. Brezis. *Analyse fonctionnelle*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1983. Théorie et applications. [Theory and applications]. 26, 104, 151

[32] F. Brezzi, L. D. Marini, S. Micheletti, P. Pietra, R. Sacco, and S. Wang. Discretization of semiconductor device problems. I. In *Handbook of numerical analysis. Vol. XIII*, Handb. Numer. Anal., XIII, pages 317–441. North-Holland, Amsterdam, 2005. 7, 41

[33] F. Brezzi, L. D. Marini, and P. Pietra. Méthodes d'éléments finis mixtes et schéma de Scharfetter-Gummel. *C. R. Acad. Sci. Paris Sér. I Math.*, 305(13) :599–604, 1987. 7, 89

[34] F. Brezzi, L. D. Marini, and P. Pietra. Two-dimensional exponential fitting and applications to drift-diffusion models. *SIAM J. Numer. Anal.*, 26(6) :1342–1355, 1989. 7, 56, 89

[35] C. J. Budd, R. Carretero-González, and R. D. Russell. Precise computations of chemotactic collapse using moving mesh methods. *J. Comput. Phys.*, 202(2) :463–487, 2005. 31, 178

[36] M. Burger, J. A. Carrillo, and M.-T. Wolfram. A mixed finite element method for nonlinear diffusion equations. *Kinet. Relat. Models*, 3(1) :59–83, 2010. 19, 31, 178

[37] M. Burger, M. Di Francesco, and Y. Dolak-Struss. The Keller-Segel model for chemotaxis with prevention of overcrowding : linear vs. nonlinear diffusion. *SIAM J. Math. Anal.*, 38(4) :1288–1315 (electronic), 2006. 25

[38] M. Burger, M. Di Francesco, J.-F. Pietschmann, and B. Schlake. Nonlinear cross-diffusion with size exclusion. *SIAM J. Math. Anal.*, 42(6) :2842–2871, 2010. 119, 123

[39] R. Bürger, , and F. Concha. Mathematical model and numerical simulation of the settling of flocculated suspensions. *Int. J. Multiphase Flow*, 24(6) :1005–1023, 1998. 18

[40] V. Calvez and J. A. Carrillo. Volume effects in the Keller-Segel model : energy estimates preventing blow-up. *J. Math. Pures Appl. (9)*, 86(2) :155–175, 2006. 25

[41] V. Calvez and L. Corrias. The parabolic-parabolic Keller-Segel model in $\mathbb{R}^2$. *Commun. Math. Sci.*, 6(2) :417–447, 2008. 24

[42] J. Carrillo. Entropy solutions for nonlinear degenerate problems. *Arch. Ration. Mech. Anal.*, 147(4) :269–361, 1999. 91

[43] J. A. Carrillo, M. Di Francesco, and M. P. Gualdani. Semidiscretization and long-time asymptotics of nonlinear diffusion equations. *Commun. Math. Sci.*, 5(suppl. 1) :21–53, 2007. 19, 123

[44] J. A. Carrillo, S. Hittmeir, and A. Jüngel. Cross diffusion and nonlinear diffusion preventing blow-up in the Keller-Segel model. To appear in *Math. Mod. Meth. Appl. Sci.*, 2012. 25, 177

[45] J. A. Carrillo, A. Jüngel, P. A. Markowich, G. Toscani, and A. Unterreiter. Entropy dissipation methods for degenerate parabolic problems and generalized Sobolev inequalities. *Monatsh. Math.*, 133(1) :1–82, 2001. 13, 14, 83, 119, 120, 205

[46] J. A. Carrillo, P. Laurençot, and J. Rosado. Fermi-Dirac-Fokker-Planck equation : well-posedness & long-time asymptotics. *J. Differential Equations*, 247(8) :2209–2234, 2009. 17, 119, 122, 138

[47] J. A. Carrillo, J. Rosado, and F. Salvarani. 1D nonlinear Fokker-Planck equations for fermions and bosons. *Appl. Math. Lett.*, 21(2) :148–154, 2008. 16, 17, 119, 122

[48] J. A. Carrillo and G. Toscani. Asymptotic L^1-decay of solutions of the porous medium equation to self-similarity. *Indiana Univ. Math. J.*, 49(1) :113–142, 2000. 15, 88, 115, 121, 136

[49] F. Cavalli, G. Naldi, G. Puppo, and M. Semplice. High-order relaxation schemes for nonlinear degenerate diffusion problems. *SIAM J. Numer. Anal.*, 45 :2098–2119, 2007. 18, 119

[50] C. Chainais-Hillairet and J. Droniou. Finite-volume schemes for noncoercive elliptic problems with Neumann boundary conditions. *IMA J. Numer. Anal.*, 31(1) :61–85, 2011. 27, 151

[51] C. Chainais-Hillairet and F. Filbet. Asymptotic behaviour of a finite-volume scheme for the transient drift-diffusion model. *IMA J. Numer. Anal.*, 27(4) :689–716, 2007. 7, 8, 9, 11, 19, 37, 41, 43, 51, 53, 54, 89, 90, 94, 111, 112, 119, 123, 133

[52] C. Chainais-Hillairet, J.-G. Liu, and Y.-J. Peng. Finite volume scheme for multi-dimensional drift-diffusion equations and convergence analysis. *M2AN Math. Model. Numer. Anal.*, 37(2) :319–338, 2003. 7, 10, 11, 53, 54, 56, 60, 61, 62, 73, 74, 89, 93, 99, 104, 106, 109, 130, 135, 180, 189

[53] C. Chainais-Hillairet and Y.-J. Peng. Convergence of a finite-volume scheme for the drift-diffusion equations in 1D. *IMA J. Numer. Anal.*, 23(1) :81–108, 2003. 7, 89, 93, 130

[54] C. Chainais-Hillairet and Y.-J. Peng. Finite volume approximation for degenerate drift-diffusion system in several space dimensions. *Math. Models Methods Appl. Sci.*, 14(3) :461–481, 2004. 7, 10, 11, 56, 57, 60, 62, 67, 89, 93, 104, 130

[55] M. Chatard. Asymptotic behavior of the Scharfetter-Gummel scheme for the drift-diffusion model. In *Finite volumes for complex applications. VI. Problems & perspectives. Volume 1, 2*, volume 4 of *Springer Proc. Math.*, pages 235–243. Springer, Heidelberg, 2011. xiii, 39

[56] Z. Chen, R. E. Ewing, Q. Jiang, and A. M. Spagnuolo. Error analysis for characteristics-based methods for degenerate parabolic problems. *SIAM J. Numer. Anal.*, 40(4) :1491–1515, 2003. 18, 119

[57] A. Chertock and A. Kurganov. A second-order positivity preserving central-upwind scheme for chemotaxis and haptotaxis models. *Numer. Math.*, 111(2) :169–205, 2008. 31, 178

[58] B. Cockburn and C.-W. Shu. The local discontinuous Galerkin method for time-dependent convection-diffusion systems. *SIAM J. Numer. Anal.*, 35(6) :2440–2463 (electronic), 1998. 18

[59] S. Cordier and E. Grenier. Quasineutral limit of an Euler-Poisson system arising from plasma physics. *Comm. Partial Differential Equations*, 25(5-6) :1099–1113, 2000. 5

[60] L. Corrias and B. Perthame. Asymptotic decay for the solutions of the parabolic-parabolic Keller-Segel chemotaxis system in critical spaces. *Math. Comput. Modelling*, 47(7-8) :755–764, 2008. 24

[61] Y. Coudière and G. Manzini. The discrete duality finite volume method for convection-diffusion problems. *SIAM J. Numer. Anal.*, 47(6) :4163–4192, 2010. 29, 153, 166

[62] Y. Coudière, T. Gallouët, and R. Herbin. Discrete Sobolev inequalities and L^p error estimates for finite volume solutions of convection diffusion equations. *M2AN, Math. Model. Numer. Anal.*, 35(4) :767–778, 2001. 27, 151

[63] Y. Coudière, J.-P. Vila, and P. Villedieu. Convergence rate of a finite volume scheme for a two-dimensional convection-diffusion problem. *M2AN Math. Model. Numer. Anal.*, 33(3) :493–516, 1999. 27, 28, 151, 166, 169

[64] R. Courant, E. Isaacson, and M. Rees. On the solution of nonlinear hyperbolic differential equations by finite differences. *Comm. Pure. Appl. Math.*, 5 :243–255, 1952. 93

[65] P. Crispel, P. Degond, and M.-H. Vignal. An asymptotic preserving scheme for the two-fluid Euler-Poisson model in the quasineutral limit. *J. Comput. Phys.*, 223(1) :208–234, 2007. 8

[66] C. Dawson. Analysis of an upwind-mixed finite element method for nonlinear contaminant transport equations. *SIAM J. Numer. Anal.*, 35(5) :1709–1724, 1998. 18

[67] P. Degond, F. Deluzet, and L. Navoret. An asymptotically stable particle-in-cell (PIC) scheme for collisionless plasma simulations near quasineutrality. *C. R. Math. Acad. Sci. Paris*, 343(9) :613–618, 2006. 8

[68] P. Degond, F. Deluzet, L. Navoret, A.-B. Sun, and M.-H. Vignal. Asymptotic-preserving particle-in-cell method for the Vlasov-Poisson system near quasineutrality. *J. Comput. Phys.*, 229(16) :5630–5652, 2010. 8

[69] S. Delcourte, K. Domelevo, and P. Omnès. Discrete-duality finite volume method for second order elliptic problems. In *Finite volumes for complex applications IV*, pages 447–458. ISTE, London, 2005. 28, 166

[70] S. Delcourte, K. Domelevo, and P. Omnes. A discrete duality finite volume approach to Hodge decomposition and div-curl problems on almost arbitrary two-dimensional meshes. *SIAM J. Numer. Anal.*, 45(3) :1142–1174, 2007. 148

[71] M. Di Francesco and J. Rosado. Fully parabolic Keller-Segel model for chemotaxis with prevention of overcrowding. *Nonlinearity*, 21(11) :2715–2730, 2008. 25

[72] D. A. Di Pietro and A. Ern. Discrete functional analysis tools for discontinuous Galerkin methods with application to the incompressible Navier-Stokes equations. *Math. Comp.*, 79(271) :1303–1330, 2010. 28, 152

[73] K. Domelevo and P. Omnes. A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids. *M2AN Math. Model. Numer. Anal.*, 39(6) :1203–1249, 2005. 28, 148, 166

[74] M. Dreher and A. Jüngel. Compact families of piecewise constant functions in $L^p(0,T;B)$. *Nonlinear Analysis : Theory, Methods & Applications*, 75(6) :3072–3077, 2012. 189

[75] J. Droniou, T. Gallouët, and R. Herbin. A finite volume scheme for a noncoercive elliptic equation with measure data. *SIAM J. Numer. Anal.*, 41(6) :1997–2031 (electronic), 2003. 27, 151

[76] L. J. Durlofsky, B. Engquist, and S. Osher. Triangle based adaptive stencils for the solution of hyperbolic conservation laws. *J. Comput. Phys.*, 98 :64–73, January 1992. 22, 126

[77] C. Ebmeyer. Error estimates for a class of degenerate parabolic equations. *SIAM J. Numer. Anal.*, 35(3) :1095–1112, 1998. 18, 119

[78] Y. Epshteyn and A. Izmirlioglu. Fully discrete analysis of a discontinuous finite element method for the Keller-Segel chemotaxis model. *J. Sci. Comput.*, 40(1-3) :211–256, 2009. 31, 178

[79] Y. Epshteyn and A. Kurganov. New interior penalty discontinuous Galerkin methods for the Keller-Segel chemotaxis model. *SIAM J. Numer. Anal.*, 47(1) :386–408, 2008. 31

[80] S. Evje and K. H. Karlsen. Viscous splitting approximation of mixed hyperbolic-parabolic convection-diffusion equations. *Numer. Math.*, 83(1) :107–137, 1999. 18, 119

[81] S. Evje and K. H. Karlsen. Discrete approximations of BV solutions to doubly nonlinear degenerate parabolic equations. *Numer. Math.*, 86(3) :377–417, 2000. 18

[82] S. Evje and K. H. Karlsen. Monotone difference approximations of BV solutions to degenerate convection-diffusion equations. *SIAM J. Numer. Anal.*, 37(6) :1838–1860 (electronic), 2000. 18

[83] R. Eymard, J. Fuhrmann, and K. Gärtner. A finite volume scheme for nonlinear parabolic equations derived from one-dimensional local Dirichlet problems. *Numer. Math.*, 102(3) :463–495, 2006. 11, 20, 91, 100, 101, 109, 110

[84] R. Eymard and T. Gallouët. H-convergence and numerical schemes for elliptic problems. *SIAM J. Numer. Anal.*, 41(2) :539–562 (electronic), 2003. 105

[85] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. In *Handbook of numerical analysis, Vol. VII*, Handb. Numer. Anal., VII, pages 713–1020. North-Holland, Amsterdam, 2000. 18, 27, 42, 57, 59, 62, 89, 92, 93, 103, 119, 130, 148, 151, 178, 179, 189

[86] R. Eymard, T. Gallouët, and R. Herbin. Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes SUSHI : a scheme using stabilization and hybrid interfaces. *IMA J. Numer. Anal.*, 30(4) :1009–1043, 2010. 27, 28, 151, 152

[87] R. Eymard, T. Gallouët, R. Herbin, and A. Michel. Convergence of a finite volume scheme for nonlinear degenerate parabolic equations. *Numer. Math.*, 92(1) :41–82, 2002. 18, 89, 119

[88] R. Eymard, D. Hilhorst, and M. Vohralík. A combined finite volume–nonconforming/mixed-hybrid finite element scheme for degenerate parabolic problems. *Numer. Math.*, 105(1) :73–131, 2006. 18, 89, 119

[89] F. Filbet. A finite volume scheme for the Patlak-Keller-Segel chemotaxis model. *Numer. Math.*, 104(4) :457–488, 2006. 27, 28, 31, 32, 148, 151, 152, 155, 177, 178, 180, 184, 185, 189, 193

[90] A. Friedman. *Partial differential equations*. Holt, Rinehart and Winston, Inc., New York, 1969. 26, 151

[91] H. Gajewski. On existence, uniqueness and asymptotic behavior of solutions of the basic equations for carrier transport in semiconductors. *Z. Angew. Math. Mech.*, 65(2) :101–108, 1985. 3

[92] H. Gajewski and K. Gärtner. On the discretization of Van Roosbroeck's equations with magnetic field. *Z. Angew. Math. Mech.*, 76(5) :247–264, 1996. 3, 4, 5, 8, 37, 41, 44, 53, 122, 203

[93] T. Gallouët, R. Herbin, and M. H. Vignal. Error estimates on the approximate finite volume solution of convection diffusion equations with general boundary conditions. *SIAM J. Numer. Anal.*, 37(6) :1935–1972 (electronic), 2000. 27, 151

[94] I. Gasser. The initial time layer problem and the quasineutral limit in a nonlinear drift diffusion model for semiconductors. *NoDEA Nonlinear Differential Equations Appl.*, 8(3) :237–249, 2001. 6, 10, 37, 57, 58, 69, 203

[95] I. Gasser, C. D. Levermore, P. A. Markowich, and C. Schmeiser. The initial time layer problem and the quasineutral limit in the semiconductor drift-diffusion model. *European J. Appl. Math.*, 12(4) :497–512, 2001. 6, 10, 37, 57, 58, 203

[96] A. Glitzky. Exponential decay of the free energy for discretized electro-reaction-diffusion systems. *Nonlinearity*, 21(9) :1989–2009, 2008. 41

[97] A. Glitzky and J. A. Griepentrog. Discrete Sobolev-Poincaré inequalities for Voronoi finite volume approximations. *SIAM J. Numer. Anal.*, 48(1) :372–391, 2010. 27, 148, 151

[98] E. Godlewski and P.-A. Raviart. *Numerical approximation of hyperbolic systems of conservation laws*, volume 118 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1996. 22, 126

[99] L. Gosse and G. Toscani. Identification of asymptotic decay to self-similarity for one-dimensional filtration equations. *SIAM J. Numer. Anal.*, 43(6) :2590–2606 (electronic), 2006. 19, 119

[100] H. Gummel. A self-consistent iterative scheme for one-dimensional steady-state transistor calculations. *IEEE Trans. Electron Dev.*, 11 :455–465, 1964. 6

[101] J. Haškovec and C. Schmeiser. Stochastic particle approximation for measure valued solutions of the 2D Keller-Segel system. *J. Stat. Phys.*, 135(1) :133–151, 2009. 31, 178

[102] J. Haškovec and C. Schmeiser. Convergence of a stochastic particle approximation for measure solutions of the 2D Keller-Segel system. *Comm. Partial Differential Equations*, 36(6) :940–960, 2011. 31, 178

[103] F. Hermeline. Une méthode de volumes finis pour les équations elliptiques du second ordre. *C. R. Acad. Sci. Paris Sér. I Math.*, 326(12) :1433–1436, 1998. 28, 166

[104] F. Hermeline. A finite volume method for the approximation of diffusion operators on distorted meshes. *J. Comput. Phys.*, 160(2) :481–499, 2000. 28, 166

[105] F. Hermeline. A finite volume method for solving Maxwell equations in inhomogeneous media on arbitrary meshes. *C. R. Math. Acad. Sci. Paris*, 339(12) :893–898, 2004. 28, 166

[106] F. Hermeline. Approximation of 2-D and 3-D diffusion operators with variable full tensor coefficients on arbitrary meshes. *Comput. Methods Appl. Mech. Engrg.*, 196(21-24) :2497–2526, 2007. 29, 166

[107] F. Hermeline. A finite volume method for approximating 3D diffusion operators on general meshes. *J. Comput. Phys.*, 228(16) :5763–5786, 2009. 29, 166

[108] M. A. Herrero, E. Medina, and J. J. L. Velázquez. Finite-time aggregation into a single point in a reaction-diffusion system. *Nonlinearity*, 10(6) :1739–1754, 1997. 24

[109] M. A. Herrero and J. J. L. Velázquez. Singularity patterns in a chemotaxis model. *Math. Ann.*, 306(3) :583–623, 1996. 24, 177

[110] T. Hillen and K. J. Painter. A user's guide to PDE models for chemotaxis. *J. Math. Biol.*, 58(1-2) :183–217, 2009. 23, 176

[111] S. Hittmeir and A. Jüngel. Cross diffusion preventing blow-up in the two-dimensional Keller-Segel model. *SIAM J. Math. Anal.*, 43(2) :997–1022, 2011. 23, 25, 32, 147, 177, 178, 181, 182, 184, 186, 192, 206

[112] D. Horstmann. From 1970 until present : the Keller-Segel model in chemotaxis and its consequences. I. *Jahresber. Deutsch. Math.-Verein.*, 105(3) :103–165, 2003. 24, 147

[113] D. Horstmann. From 1970 until present : the Keller-Segel model in chemotaxis and its consequences. II. *Jahresber. Deutsch. Math.-Verein.*, 106(2) :51–69, 2004. 24

[114] A. M. Il'in. Differencing scheme for a differential equation with a small parameter affecting the highest derivative. *Mathematical Notes*, 6 :596–602, 1969. 10.1007/BF01093706. 6, 37, 41, 44, 94, 130

[115] W. Jäger and J. Kačur. Solution of porous medium type systems by linear approximation schemes. *Numer. Math.*, 60(3) :407–427, 1991. 18, 119

[116] W. Jäger and S. Luckhaus. On explosions of solutions to a system of partial differential equations modelling chemotaxis. *Trans. Amer. Math. Soc.*, 329(2) :819–824, 1992. 24, 147, 177

[117] S. Jin. Efficient Asymptotic-Preserving (AP) Schemes For Some Multiscale Kinetic Equations. *SIAM J. Sci. Comput.*, 21(2) :441–454, 1999. 7

[118] A. Jüngel. Numerical approximation of a drift-diffusion model for semiconductors with nonlinear diffusion. *Z. Angew. Math. Mech.*, 75(10) :783–799, 1995. 7, 11, 20, 56, 89, 91, 95

[119] A. Jüngel. Qualitative behavior of solutions of a degenerate nonlinear drift-diffusion model for semiconductors. *Math. Models Methods Appl. Sci.*, 5(4) :497–518, 1995. 16, 41, 44, 87, 88, 122, 203

[120] A. Jüngel. *Quasi-hydrodynamic semiconductor equations*. Progress in Nonlinear Differential Equations and their Applications, 41. Birkhäuser Verlag, Basel, 2001. 2, 56

[121] A. Jüngel and Y.-J. Peng. A hierarchy of hydrodynamic models for plasmas. Quasi-neutral limits in the drift-diffusion equations. *Asymptot. Anal.*, 28(1) :49–73, 2001. 6, 10, 37, 57, 58, 62, 203

[122] A. Jüngel and P. Pietra. A discretization scheme for a quasi-hydrodynamic semiconductor model. *Math. Models Methods Appl. Sci.*, 7(7) :935–955, 1997. 7, 11, 20, 56, 89, 91, 95

[123] J. Kačur. Solution of degenerate convection-diffusion problems by the method of characteristics. *SIAM J. Numer. Anal.*, 39(3) :858–879, 2002. 18, 119

[124] G. Kaniadakis and P. Quarati. Kinetic equation for classical particles obeying an exclusion principle. *Phys. Rev. E*, 48 :4263–4270, Dec 1993. 16

[125] G. Kaniadakis and P. Quarati. Classical model of bosons and fermions. *Phys. Rev. E*, 49 :5103–5110, Jun 1994. 16

[126] K. H. Karlsen, N. H. Risebro, and J. D. Towers. Upwind difference approximations for degenerate parabolic convection–diffusion equations with a discontinuous coefficient. *IMA J. Numer. Anal.*, 22(4) :623, 2002. 119

[127] E. F. Keller and L. A. Segel. Initiation of slime mold aggregation viewed as an instability. *J. Theor. Biol.*, 26(3) :399–415, 1970. 23, 176

[128] R. Kowalczyk. Preventing blow-up in a chemotaxis model. *J. Math. Anal. Appl.*, 305(2) :566–588, 2005. 25

[129] S. Krell. Stabilized DDFV schemes for Stokes problem with variable viscosity on general 2D meshes. *Numer. Methods Partial Differential Equations*, 27(6) :1666–1706, 2011. 29, 166

[130] A. Kurganov and E. Tadmor. New high-resolution central schemes for nonlinear conservation laws and convection-diffusion equations. *J. Comput. Phys.*, 160(1) :241–282, 2000. 18, 119, 143

[131] A. Lasis and E. Süli. Poincaré-type inequalities for broken Sobolev spaces. Technical Report 03/10. Technical report, Oxford University Computing Laboratory, Oxford, England, 2003. 28, 152

[132] R. D. Lazarov, I. D. Mishev, and P. S. Vassilevski. Finite volume methods for convection-diffusion problems. *SIAM J. Numer. Anal.*, 33(1) :31–55, 1996. 20, 41, 44, 83, 94, 130

[133] A. H. Le and P. Omnès. Discrete Poincaré inequalities for arbitrary meshes in the discrete duality finite volume context. Preprint 2012. 172

[134] E. Lieb and M. Loss. *Analysis*, volume 14 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1997. 156

[135] Y. Liu, C.-W. Shu, and M. Zhang. High order finite difference WENO schemes for nonlinear degenerate parabolic equations. *SIAM J. Sci. Comput.*, 33(2) :939–965, 2011. 18, 119, 143

[136] P. A. Markowich. *The stationary semiconductor device equations*. Computational Microelectronics. Springer-Verlag, Vienna, 1986. 4, 87

[137] P. A. Markowich, C. A. Ringhofer, and C. Schmeiser. *Semiconductor equations*. Springer-Verlag, Vienna, 1990. 2, 3, 4, 56, 87

[138] P. A. Markowich and A. Unterreiter. Vacuum solutions of a stationary drift-diffusion model. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)*, 20(3) :371–386, 1993. 15, 87, 121

[139] A. Marrocco. Numerical simulation of chemotactic bacteria aggregation via mixed finite elements. *ESAIM : M2AN*, 37(04) :617–630, 2003. 31, 178

[140] M. S. Mock. Asymptotic behavior of solutions of transport equations for semiconductor devices. *J. Math. Anal. Appl.*, 49 :215–225, 1975. 3

[141] T. Nagai. Blow-up of radially symmetric solutions to a chemotaxis system. *Adv. Math. Sci. Appl.*, 5(2) :581–601, 1995. 194

[142] T. Nagai. Blowup of nonradial solutions to parabolic-elliptic systems modeling chemotaxis in two-dimensional domains. *J. Inequal. Appl.*, 6 :37–55, 2001. 24, 176, 193

[143] T. Nagai, T. Senba, and K. Yoshida. Application of the Trudinger-Moser inequality to a parabolic system of chemotaxis. *Funkcial. Ekvac.*, 40 :411–433, 1997. 24, 176

[144] L. Nirenberg. On elliptic partial differential equations. *Ann. Scuola Norm. Sup. Pisa (3)*, 13 :115–162, 1959. 26, 151

[145] R. H. Nochetto, A. Schmidt, and C. Verdi. A posteriori error estimation and adaptivity for degenerate parabolic problems. *Math. Comp.*, 69(229) :1–24, 2000. 18, 119

[146] R. H. Nochetto and C. Verdi. Approximation of degenerate parabolic problems using a numerical integration. *SIAM J. Numer. Anal.*, 25(4) :784–814, 1988. 18, 119

[147] M. Ohlberger. A posteriori error estimates for vertex centered finite volume approximations of convection-diffusion-reaction equations. *M2AN Math. Model. Numer. Anal.*, 35(2) :355–387, 2001. 18, 119

[148] K. Osaki and A. Yagi. Finite dimensional attractor for one-dimensional Keller-Segel equations. *Funkcial. Ekvac.*, 44 :441–469, 2001. 24

[149] C. S. Patlak. Random walk with persistence and external bias. *Bull. Math. Biology*, 15 :311–338, 1953. 10.1007/BF02476407. 23, 176

[150] B. Perthame. PDE models for chemotactic movements : parabolic, hyperbolic and kinetic. *Appl. Math.*, 49(6) :539–564, 2004. 194

[151] B. Perthame. *Transport equations in biology*. Frontiers in Mathematics. Birkhäuser Verlag, Basel, 2007. 23, 176

[152] I. S. Pop and W.-A. Yong. A numerical approach to degenerate parabolic equations. *Numer. Math.*, 92(2) :357–381, 2002. 18, 119

[153] A. Prohl and M. Schmuck. Convergent discretizations for the Nernst-Planck-Poisson system. *Numer. Math.*, 111(4) :591–630, 2009. 45

[154] C. Ringhofer. An asymptotic anaylsis of a transient $p-n$-junction model. *SIAM J. Appl. Math.*, 47(3) :624–642, 1987. 6

[155] N. Saito. Conservative upwind finite-element method for a simplified Keller-Segel system modelling chemotaxis. *IMA J. Numer. Anal.*, 27(2) :332–365, 2007. 31, 178

[156] N. Saito. Error analysis of a conservative finite-element approximation for the Keller-Segel system of chemotaxis. *Commun. Pure Appl. Anal.*, 11(1) :339–364, 2012. 31, 178

[157] N. Saito and T. Suzuki. Notes on finite difference schemes to a parabolic-elliptic system modelling chemotaxis. *Appl. Math. Comput.*, 171(1) :72–90, 2005. 31, 178

[158] D. L. Scharfetter and H. K. Gummel. Large signal analysis of a silicon Read diode. *IEEE Trans. Elec. Dev.*, 16 :64–77, 1969. 6, 37, 41, 44, 89, 94, 130

[159] R. Strehl, A. Sokolov, D. Kuzmin, and S. Turek. A flux-corrected finite element method for chemotaxis problems. *Comput. Methods Appl. Math.*, 10(2) :219–232, 2010. 31, 178

[160] G. Toscani. Finite time blow up in Kaniadakis-Quarati model of Bose-Einstein particles. *Comm. Partial Differential Equations*, 37(1) :77–87, 2012. 17, 119, 122, 123

[161] R. Tyson, L. G. Stern, and R. J. LeVeque. Fractional step methods applied to a chemotaxis model. *J. Math. Biol.*, 41(5) :455–475, 2000. 31, 178

[162] W. van Roosbroeck. Theory of flow of electrons and holes in germanium and other semiconductors. *Bell Syst. Techn. J.*, 29 :560–607, 1950. 2

[163] Martin Vohralík. On the discrete Poincaré-Friedrichs inequalities for nonconforming approximations of the Sobolev space H^1. *Numer. Funct. Anal. Optim.*, 26(7-8) :925–952, 2005. 28, 152

[164] M. Winkler. Chemotaxis with logistic source : very weak global solutions and their boundedness properties. *J. Math. Anal. Appl.*, 348(2) :708–729, 2008. 25

[165] M. Winkler. Aggregation vs. global diffusive behavior in the higher-dimensional Keller-Segel model. *J. Differential Equations*, 248(12) :2889–2905, 2010. 24

[166] Q. Zhang and Z.-L. Wu. Numerical simulation for porous medium equation by local discontinuous Galerkin finite element method. *J. Sci. Comput.*, 38(2) :127–148, 2009. 18, 119

[167] W. P. Ziemer. *Weakly differentiable functions*. Springer-Verlag New York, Inc., New York, NY, USA, 1989. 29, 155

Résumé

Cette thèse est dédiée au développement et à l'analyse de schémas numériques de type volumes finis pour des équations de convection-diffusion, qui apparaissent notamment dans des modèles issus de la physique ou de la biologie. Nous nous intéressons plus particulièrement à la préservation de comportements asymptotiques au niveau discret. Ce travail s'articule en trois parties, composées chacune de deux chapitres.

Dans la première partie, nous considérons la discrétisation du système de dérive-diffusion linéaire pour les semi-conducteurs par le schéma de Scharfetter-Gummel implicite en temps. Nous nous intéressons à la préservation par ce schéma de deux types d'asymptotiques : l'asymptotique en temps long et la limite quasi-neutre. Nous démontrons des estimations d'énergie–dissipation d'énergie discrètes qui permettent de prouver d'une part la convergence en temps long de la solution approchée vers une approximation de l'équilibre thermique, d'autre part la stabilité à la limite quasi-neutre du schéma.

Dans la deuxième partie, nous nous intéressons à des schémas volumes finis préservant l'asymptotique en temps long dans un cadre plus général. Plus précisément, nous considérons des équations de type convection-diffusion non linéaires qui apparaissent dans plusieurs contextes physiques : équations des milieux poreux, système de dérive-diffusion pour les semi-conducteurs... Nous proposons deux discrétisations en espace permettant de préserver le comportement en temps long des solutions approchées. Dans un premier temps, nous étendons la définition du flux de Scharfetter-Gummel pour une diffusion non linéaire. Ce schéma fournit des résultats numériques satisfaisants si la diffusion ne dégénère pas. Dans un second temps, nous proposons une discrétisation dans laquelle nous prenons en compte ensemble les termes de convection et de diffusion, en réécrivant le flux sous la forme d'un flux d'advection. Le flux numérique est défini de telle sorte que les états d'équilibre soient préservés, et nous utilisons une méthode de limiteurs de pente pour obtenir un schéma précis à l'ordre deux en espace, même dans le cas dégénéré.

Enfin, la troisième et dernière partie est consacrée à l'étude d'un schéma numérique pour un modèle de chimiotactisme avec diffusion croisée pour lequel les solutions n'explosent pas en temps fini, quelles que soient les données initiales. L'étude de la convergence du schéma repose sur une estimation d'entropie discrète nécessitant l'utilisation de versions discrètes d'inégalités fonctionnelles telles que les inégalités de Poincaré-Sobolev et de Gagliardo-Nirenberg-Sobolev. La démonstration de ces inégalités fait l'objet d'un chapitre indépendant dans lequel nous proposons leur étude dans un contexte assez général, incluant notamment le cas de conditions aux limites mixtes et une généralisation au cadre des schémas DDFV.

Abstract

This dissertation is dedicated to the development and analysis of finite volume numerical schemes for convection-diffusion equations, which notably occur in models arising from physics and biology. We are more particularly interested in preserving asymptotic behavior at the discrete level. This dissertation is composed of three parts, each one including two chapters.

In the first part, we consider the discretization of the linear drift-diffusion system for semiconductors with the implicit Scharfetter-Gummel scheme. We focus on preserving two kinds of asymptotics with this scheme : the long-time asymptotic and the quasineutral limit. We show discrete energy–energy dissipation estimates which constitute the main point to prove first the large time convergence of the approximate solution to an approximation of the thermal equilibrium, and then the stability at the quasineutral limit.

In the second part, we are interested in designing finite volume schemes which preserve the long time behavior in a more general framework. More precisely, we consider nonlinear convection-diffusion equations arising in various physical models : porous media equation, drift-diffusion system for semiconductors... We propose two spatial discretizations which preserve the long time behavior of the approximate solutions. We first generalize the Scharfetter-Gummel flux for a nonlinear diffusion. This scheme provides satisfying numerical results if the diffusion term does not degenerate. Then we propose a discretization which takes into account together the convection and diffusion terms by rewriting the flux as an advective flux. The numerical flux is then defined in such a way that equilibrium states are preserved, and we use a slope limiters method so as to obtain second order space accuracy, even in the degenerate case.

Finally, the third part is devoted to the study of a numerical scheme for a chemotaxis model with cross diffusion, for which the solutions do not blow up in finite time, even for large initial data. The proof of convergence is based on a discrete entropy estimate which requires the use of discrete functional inequalities such as Poincaré-Sobolev and Gagliardo-Nirenberg-Sobolev inequalities. The demonstration of these inequalities is the subject of an independent chapter in which we propose a study in quite a general framework, including mixed boundary conditions and generalization to DDFV schemes.